Constrained Quadratic Optimization: Theory and Application for Wireless Communication Systems

by

Erik Hons

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Master of Applied Science

in

Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2001

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the University of Waterloo to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

The suitability of general constrained quadratic optimization as a modeling and solution tool for wireless communication is discussed. It is found to be appropriate due to the many quadratic entities present in wireless communication models. It is posited that decisions which affect the *minimum mean squared error* at the receiver may achieve improved performance if those decisions are *energy constrained*. That theory is tested by applying a specific type of constrained quadratic optimization problem to the forward link of a cellular DS-CDMA system. Through simulation it is shown that when channel coding is used, energy constrained methods typically outperform unconstrained methods. Furthermore, a new energy-constrained method is presented which significantly outperforms a similar published method in several different environments.

Acknowledgements

I would like to acknowledge my supervisor Dr. A. K. Khandani for his guidance throughout my research. His confidence in me, from the very beginning, made all of this possible.

Contents

List of Figures

Chapter 1

Introduction

Many engineering problems can be stated as constrained numerical optimization problems. That is, some performance metric is being optimized with respect to physical or design limits. Whether this is useful depends on the class of optimization problem. Considerable insight can be gained if the problem class has a large body of existing results, proofs, and solution algorithms. If, for example, a problem can be modeled within the class of linear programming, then a solution algorithm – the simplex method – can be applied immediately along with many results concerning termination, degeneracy, etc.

Consider the conceptualized model of a digital communication system pictured in Figure 1.1. In this model, a digital source (possibly generated from a continuous process by an analog to digital converter (ADC)) feeds data symbols to a source coder which removes redundancy to achieve efficient representation of the source. The source coded symbols are then fed to a channel coder which adds controlled redundancy to improve error performance. The channel coded symbols are then

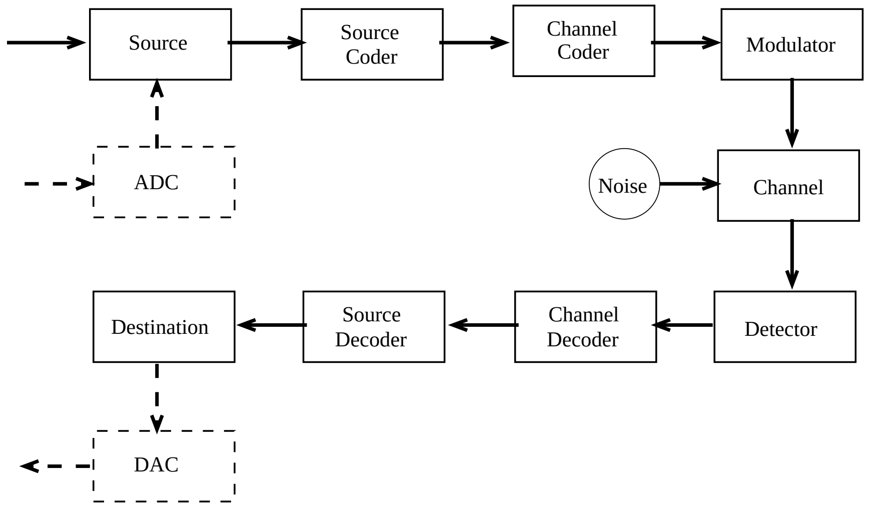

Figure 1.1: Conceptualized System Model

modulated and emitted as signals through a noisy channel. The received signals are first converted into decision statistics by a detector and then fed to channel and source decoders which attempt to recover the original data symbols as accurately as possible (a digital to analog converter (DAC) is applied if necessary).

A common goal of such systems is to achieve the optimum end-to-end error rate for a given signal to noise ratio (SNR). The receiver achieves this by using metrics, such as *minimum mean-square error* (MMSE), to select the most likely member of a signal constellation based on channel observations. In practical versions of this model, either bandwidth, transmitted energy, or both are constrained. When the signal constellation is designed then, a trade-off must be made between optimum receiver metrics, and the practical system constraints.

Because the objective is finding the optimal value for a system metric, and because this optimization is constrained by physical limits, practical communications systems are, fundamentally, constrained optimization problems. Typically, communication systems are divided into sections which are optimized independently. Thus, one may talk of optimizing the receiver structure of a multiuser system, or minimizing the out-of-band emissions of a particular signal constellation. Regardless, in practical systems, the problems are always optimization problems in nature.

The above communications model contains several quadratic entities. *Distortion* can occur through source quantization or using fewer channel symbols than source symbols. *Power* is consumed and radiated by the transmitter. Minimum mean-square *error* is the criterion for maximum likelihood detection. Underlying each of these quadratic metrics is the concept of squared Euclidean distance. This observation is important because Euclidean distance is a *convex function*. That is, for any two points x and y, the graph of the Euclidean distance function $\| \cdot \|$ lies below the line segment joining $(x, \|x\|)$ and $(y, \|y\|)$ or, more formally,

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \quad \|\alpha \mathbf{x} + (1-\alpha)\mathbf{y}\| \leq \alpha \|\mathbf{x}\| + (1-\alpha)\|\mathbf{y}\|, \quad 0 \leq \alpha \leq 1 \tag{1.1}$$

Additionally, when Euclidean distance is constrained the result is a *convex set*. That is, the set of solutions to $\|\mathbf{x}\| \leq r$ forms a convex set where a set $\mathbf{S} \in \mathbb{R}^n$ is called convex if the straight line joining any two points in $\mathbf{S}$ is contained entirely in $\mathbf{S}$.

The convexity of these metrics has important implications for the solvability of their respective problems. Convexity implies the problem has a benign structure in several respects. For example, convex optimization problems, whose objective metrics and constraint sets are both convex, have the property that all locally optimum solutions are also globally optimum. Features like this can greatly reduce the complexity of solution algorithms.

Because of the abundance of quadratic entities and the generally convex nature of the performance metrics, it makes sense to consider quadratic optimization as a candidate both for modeling and solving communications problems. Several types of quadratic optimization are of particular importance. When the constraints restrict the solution to be binary, the result is a quadratic 0-1 program which can be difficult to solve for large problems. If the constraints are linear, the problem is referred to as a quadratic program. In many cases, quadratic programs can be as easy to solve as a standard linear program. In communications, quadratic entities will often be related in some way so that a quadratic objective will be subject to quadratic constraints. These problems can be solved efficiently in specialized cases only, one of which, fortunately, is a convex problem.

This thesis will focus on applications of quadratic optimization to problems in communication engineering. First, an introduction to the general theory of numerical optimization will be given in chapter 2, followed by selected results and algorithms for quadratic optimization. Those techniques will be applied to efficiently implement a new method for improving a special type of down-link in a cellular system which will be proposed in chapter 3. Numerical results presented in that chapter will will show that the proposed method performs particularly well when forward error correction is used. Finally chapter 4 will make some concluding remarks.

Chapter 2

Quadratic Optimization in Communications

2.1 General Constrained Optimization

For $\mathbf{x} \in \mathbb{R}^n$, the general constrained optimization problem can be stated as[1].

$$\min_{\mathbf{x} \in \Omega} f(\mathbf{x}) \tag{2.1}$$

This specifies the minimization of an *objective function* $f() : \mathbb{R}^n \rightarrow \mathbb{R}$ over a *feasible set* Ω, where Ω is defined as

[1] Nocedal and Wright give a detailed introduction to numerical optimization in [1]. That introduction is summarized briefly in this section

$$\Omega = \{\mathbf{x} \,|\, c_i(\mathbf{x}) = 0,\, i \in \mathcal{E};\, c_i(\mathbf{x}) \geq 0,\, i \in \mathcal{I}\} \tag{2.2}$$

The functions $c_i() : \mathbb{R}^n \rightarrow \mathbb{R}$ are referred to collectively as the *constraints* and are either equalities or inequalities. The sets $\mathcal{E}$ and $\mathcal{I}$ are finite non-overlapping index sets which indicate the type of constraint each $c_i()$ specifies. If $c_i(\mathbf{x}) = 0$ when $i \in \mathcal{E}$, or $c_i(\mathbf{x}) \geq 0$ when $i \in \mathcal{I}$, the constraint is said to be *satisfied.* The feasible set, then, is the set of $\mathbf{x}$ for which all of the constraints are satisfied. An arbitrary point $\mathbf{x}$ is said to be *feasible* if $\mathbf{x} \in \Omega$. Such a point will satisfy some of the inequality constraints with $c_i(\mathbf{x}) = 0$ and others with $c_i(\mathbf{x}) > 0$. Any constraint satisfied with strict equality at $\mathbf{x}$ is called *active* while any constraint satisfied with strict inequality is called *inactive.* This last definition is important for characterizing solutions.

The problem stated in (2.1) is a search for any point $\mathbf{x}^*$ for which $f(\mathbf{x}^*) \leq f(\mathbf{x}),\ \forall \mathbf{x} \in \Omega$. Such a point is said to be a *global optimum* point. Several such points may exist, but they all must share one optimal value. During the search, *local* optimum points may discovered. At a local optimum $\mathbf{x}$, there exists an open set $\mathcal{N} \subseteq \Omega$ containing $\mathbf{x}^*$ (referred to as a *neighbourhood* of $\mathbf{x}^*$) such that $f(\mathbf{x}^*) \leq f(\mathbf{x}),\ \forall \mathbf{x} \in \mathcal{N}$.

Searching for the global optimum solution involves searching for the best local optimum. Local solutions have distinguishing characteristics which can aid in this search. To introduce the topic briefly, there are two types of local solutions: those at which one or more constraints are active, and those at which no constraints are active. Problems for which $\mathcal{E} = \varnothing$ permit solutions which have no active constraints. These solutions are strictly inside the feasible set, and correspond to local unconstrained minimization of the objective function. Thus, for a local optimum strictly inside the feasible set, the gradient to the objective function must be zero. More often, optimal points will exist at the limits of the feasible set. In this case, the gradient to the combined set of active constraints must be parallel to the objective gradient. These characteristics are summarized in the well known *Karush-Kuhn-Tucker* (KKT) conditions which are the first-order necessary conditions for a point $\mathbf{x}^*$ to be a local optimum. Using the *Lagrangian function*

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda_1 c_1(\mathbf{x}) - \cdots - \lambda_n c_n(\mathbf{x}) \tag{2.3}$$

where λ is a vector of real-valued *Lagrange multipliers*, the KKT conditions state that if $\mathbf{x}^*$ is a local optimum of (2.1) then there exits a λ^* such that

$$\begin{aligned}
\nabla_x \mathcal{L}(\mathbf{x}^*, \lambda^*) &= 0 && \text{(2.4a)}\\
c_i(\mathbf{x}^*) &= 0, \qquad \forall i \in \mathcal{E}, && \text{(2.4b)}\\
c_i(\mathbf{x}^*) &\geq 0, \qquad \forall i \in \mathcal{I}, && \text{(2.4c)}\\
\lambda_i^* &\geq 0, \qquad \forall i \in \mathcal{I}, && \text{(2.4d)}\\
\lambda_i^* c_i(\mathbf{x}^*) &= 0, \qquad \forall i \in \mathcal{E} \cup \mathcal{I}. && \text{(2.4e)}
\end{aligned}$$

For the KKT conditions to hold, it is additionally required that the set of gradients to the active constraints at $\mathbf{x}^*$ be linearly independent. Note that linear independence fails whenever an active constraint gradient is zero.

Equation (2.4e) is called the *complementarity condition*. Whenever a Lagrange multiplier is positive, the complementarity condition requires that the corresponding constraint is active. Conversely, if the constraint is not active, the corresponding Lagrange multiplier must be zero. *Strict complementarity* holds if exactly one of λ_i^* and $c_i(\mathbf{x}^*)$ are zero. When it does, the optimal λ^* is unique. When it does not, there may be many λ^* for which the KKT conditions hold.

The KKT conditions tell us how the objective function and constraint gradients are related at $\mathbf{x}^*$. They require that all the first derivatives of $f()$ be either positive or zero. They cannot tell us, however, whether the zero derivatives of $f()$ are due to positive or negative curvature at $\mathbf{x}^*$. Additional second order conditions are needed to establish whether $\mathbf{x}^*$ is, indeed, optimal. If $\mathbf{x}^*$ is a local optimum of (2.1) for which λ^* satisfies the KKT conditions, and $\mathbf{d}$ is any vector of movement from $\mathbf{x}^*$ that maintains feasibility, then the second order necessary condition is

$$\mathbf{d}^t \nabla_{xx} \mathcal{L}(\mathbf{x}^*, \lambda^*) \mathbf{d} \geq 0 \tag{2.5}$$

These definitions are enough to perform the analysis in the next section. The reader is referred to [1] for a thorough development of general constrained optimization including proofs for the first and second order conditions. We continue by introducing constrained quadratic optimization.

2.2 Constrained Quadratic Optimization

We consider here the quadratic specialization of (2.1) shown below with the objective function $q()$ in matrix form

$$\begin{aligned} \min_{\mathbf{x}} \quad & q(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^t\mathbf{Q}\mathbf{x} + \mathbf{x}^t\mathbf{c} \\ \text{subject to} \quad & \begin{cases} c_i(\mathbf{x}) = 0, i \in \mathcal{E} \\ c_i(\mathbf{x}) \geq 0, i \in \mathcal{I} \end{cases} \end{aligned} \tag{2.6}$$

where $\mathbf{x}^t$ denotes the transpose of vector $\mathbf{x}$.

Several specializations of (2.6) are particularly useful. If the constraints $c_i()$ are linear inequalities, the feasible set is a convex polytope, and we obtain the problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & q(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^t\mathbf{Q}\mathbf{x} + \mathbf{x}^t\mathbf{c} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} + \mathbf{b} \leq 0 \end{aligned} \tag{2.7}$$

which is referred to as a *quadratic program* (QP). Quadratic programs are becoming increasingly popular as alternatives to linear programs in business and economic models.

In general, quadratic programs are difficult to solve. The difficulty in solving a QP depends on the structure of $q()$. If the Hessian $\nabla^2 q(\mathbf{x}) = Q$ is positive semidefinite[2], then $q()$ is a convex function and (2.7) is called a *convex* QP. In this case,

[2] We say a symmetric matrix $\mathbf{A}$ is *positive definite* if there exists a positive constant α such that $\mathbf{x}^t\mathbf{A}\mathbf{x} \geq \alpha\|\mathbf{x}\|^2,$ for all $\mathbf{x} \in \mathbb{R}^n$. This is equivalent to saying $\mathbf{A}$ has only positive eigenvalues.

(2.7) is not significantly harder to solve than a linear program [2]. Conversely, the non-convex case is classified as NP-complete[3] [3], and is thus intractable for large problems.

What is interesting is when two quadratic entities are related. Consider the following:

$$\begin{array}{rl} \min\limits_{\mathbf{x}} & q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^t\mathbf{Q}\mathbf{x} + \mathbf{x}^t\mathbf{c} \\ \text{subject to} & \frac{1}{2}\mathbf{x}^t\mathbf{R}\mathbf{x} + \mathbf{x}^t\mathbf{d} \leq r \end{array} \tag{2.8}$$

This formulation *caps* the value of one quadratic transformation of $\mathbf{x}$ and then minimizes another quadratic transformation subject to that constraint. Alternatively, we can *fix* the value of the constraint transformation by changing the inequality in the constraint to equality. If we assume that the constraint function is convex, we obtain minimization over an ellipsoid or an ellipsoid surface in the equality constrained case. Note that when (2.8) has a convex constraint, we can apply an invertible linear transformation to the problem that converts the ellipsoid into a sphere of radius $\sqrt{r}$. This results in

We say $\mathbf{A}$ is positive *semi*definite if the above holds for $\alpha = 0$ or equivalently if $\mathbf{A}$ has only non-negative eigenvalues.

[3] Problems classified as NP-complete have no known sub-exponential solution algorithms. Moreover, an NP-complete problem is, by definition, demonstrably as difficult to solve as any other NP-complete problem.

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & q(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^t\mathbf{Q}\mathbf{x} + \mathbf{x}^t\mathbf{c} \\
\text{subject to} \quad & \|\mathbf{x}\| \leq r
\end{aligned}
\tag{2.9}
$$

For search algorithms, a sphere is the simplest constraint to consider because surface gradients can be calculated easily (the Hessian for a sphere is simply the identity matrix **I**). In [4], this formulation is used as a sub-problem in a solution algorithm for (2.7) when the objective function is non-convex. In fact, (2.9) occurs frequently as a step in so-called interior-point algorithms for general constrained optimization[4].

When both $q()$ and the constraint are convex, (2.7), (2.8), and (2.9) can have particularly simple solutions. Below, we present an algorithm due to [4], which solves (2.9) under this convexity condition. The algorithm solves the inequality constrained problem, minimization over a sphere. Later, we show in this thesis how the algorithm can be extended to solve the equality constrained problem despite the non-convex constraint set. We begin with an intuitive justification for the KKT conditions.

First, consider the geometric situation depicted in Figure 2.1. This represents the structure of the the convex form of (2.9) in two dimensions with the stationary point of the constraint occurring inside the feasible set. Clearly the global optimum solution is x_1, the stationary point of $q()$. At this point, the single inequality constraint is satisfied but inactive. To see that x_1 is the optimum, consider the non-

[4] For an example of this application, see [5]

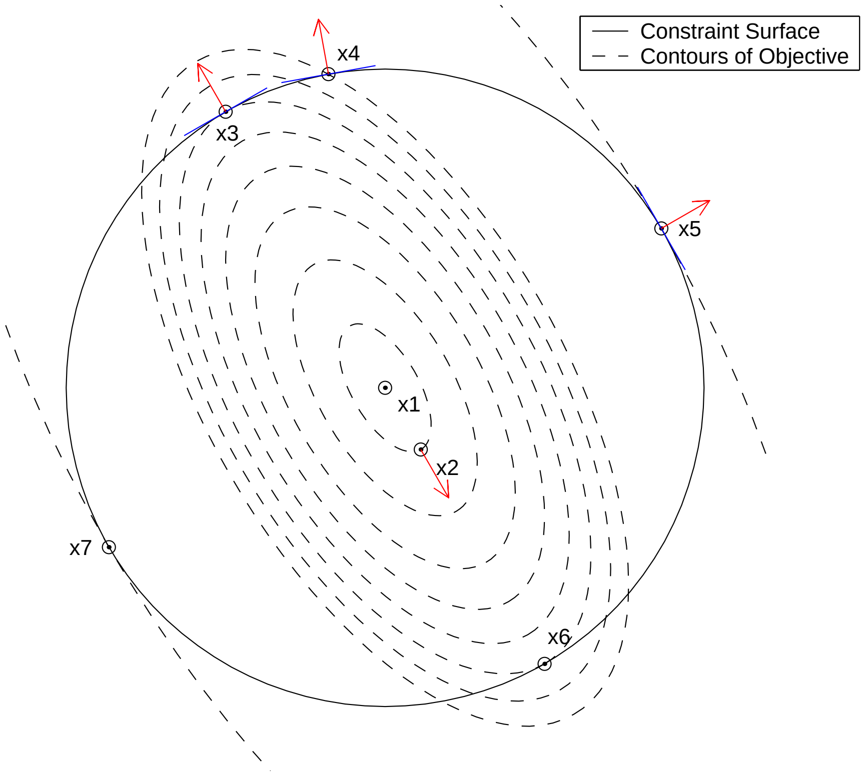

Figure 2.1: Geometrical depiction of convex quadratic optimization: $\lambda^* = 0$

optimum point x_2. At x_2, movement along the line toward x_1 not only maintains feasibility but also decreases the value of $q()$. Therefore, no neighbourhood of x_2 exists for which x_2 is a local optimum.

If we restrict our attention to points on the constraint surface, we obtain conditions similar to those at either x_3 or x_4. Consider first, point x_4. Here, the objective gradient $\nabla q(x_4)$ is not parallel to the constraint gradient $\nabla c(x_4)$. Therefore, a movement in one direction along the circumference of the circle will both maintain feasibility and reduce the objective value. Contrast this with point x_3 where the gradients are parallel. Here, any movement along the circumference results in an increase in $q()$. In both cases, however, movement toward x_1 is an improvement, so neither point can be optimal.

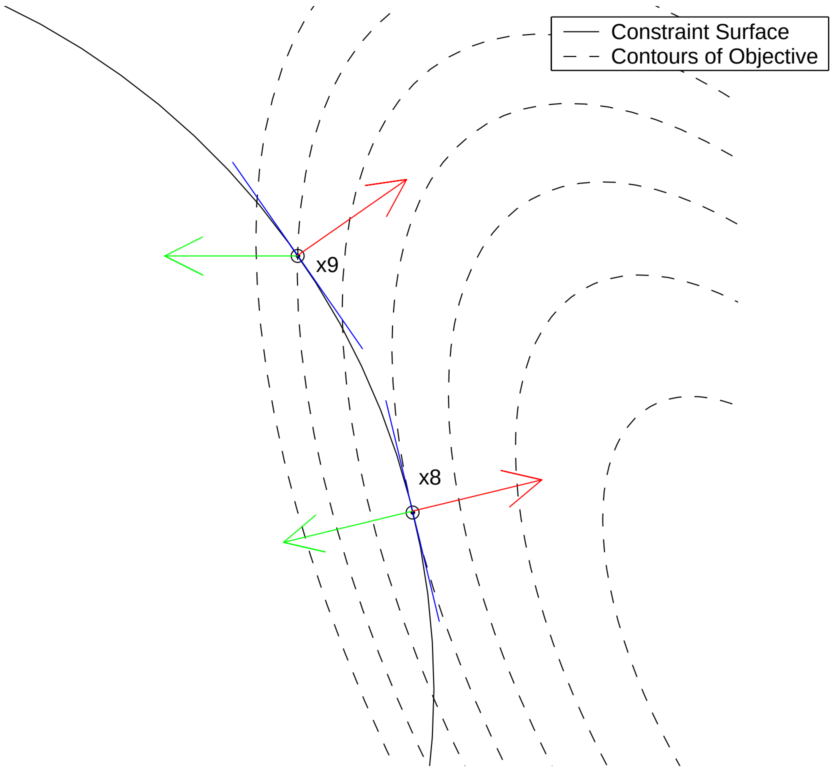

Figure 2.2: Geometrical depiction of convex quadratic optimization: $\lambda^* > 0$

Contrast this with the situation in Figure 2.2. Here the stationary point of $q()$ is *not* feasible. By (2.4a), the non-zero gradient of $q()$ requires that the corresponding Lagrange multiplier must be nonzero at an optimal point, so that (2.4e) requires the constraint to be active. Two points which match this situation are x_8 and x_9. At x_9, the gradients are not parallel and a direction of movement along the surface improves $q()$. While at x_8, the gradients are parallel and no such direction exists. Therefore, a neighbourhood of x_8 exists for which x_8 is the optimum. The problem's convexity implies that the local optimum x_8 is also the global optimum.

Clearly, the solutions to (2.9) are strongly characterized by the objective and constraint gradients, a fact which can be exploited to perform the search. At a point $\mathbf{x}$, the gradient to the objective function is given by $\mathbf{Qx} + \mathbf{c}$ while the gradient to the constraint $c(\mathbf{x}) = r - \|\mathbf{x}\|$ is given by $-2I\mathbf{x}$ where I is the identity matrix. If the KKT conditions for local optimum $\mathbf{x}^*$ are satisfied with Lagrange multiplier[5] λ^* then the following hold:

$$(\mathbf{Q} + \lambda^*\mathbf{I})\mathbf{x}^* + \mathbf{c} = 0 \tag{2.10a}$$

$$r - \|\mathbf{x}^*\| \geq 0 \tag{2.10c}$$

$$\lambda^* \geq 0 \tag{2.10d}$$

$$\lambda^*(r - \|\mathbf{x}^*\|) = 0 \tag{2.10e}$$

where the index letters indicate correspondence with the items in (2.4a)-(2.4e). The factor 2 has been absorbed into λ^* for notational convenience without affecting the final solution.

The second order necessary condition for this problem can be stated as

$$\mathbf{d}^t(\mathbf{Q} + \lambda^*\mathbf{I})\mathbf{d} \geq 0, \quad \forall \mathbf{d} \in \{\mathbf{w} | \mathbf{w}^t\mathbf{x}^* = 0\} \tag{2.11}$$

Because we have assumed $q()$ is convex, and because the KKT conditions restrict

[5] For a more complete discussion of Lagrange multipliers and optimality, see [6]

λ^* to be positive, $(\mathbf{Q} + \lambda^*\mathbf{I})$ is always positive definite. Clearly then, (2.11) will be satisfied for any point. Moreover, because $c()$ is a sphere constraint, its gradient is nonzero at all points other than the origin. This allows us to conclude that for any local optimum point other than the origin, the KKT conditions are satisfied with a unique λ^*. Moreover, from the convexity of $q()$ and $c()$, we can conclude that all local optimums are global optimums and thus the λ^* is itself unique.

At this point it will be helpful to have a convenient notation for the eigenvalues of a matrix. For matrix $\mathbf{A}$, we use $\psi_i(\mathbf{A})$ to denote a particular eigenvalue, $\psi(\mathbf{A})$ to denote *any* eigenvalue, and $\underline{\psi}(\mathbf{A})$ to denote the *least* eigenvalue of $\mathbf{A}$[6]. Moreover, we assume that the eigenvalues are labeled by their magnitude so that $\underline{\psi}(\mathbf{A}) = \psi_1(\mathbf{A}) \leq \cdots \leq \psi_n(\mathbf{A})$.

The range of possible values for λ^* is already lower bounded by zero. We can obtain an upper bound for λ by rearranging and taking the norm of each side of (2.10a). As shown in [4], if we let $\underline{\psi}(\mathbf{Q})$ represent the least eigenvalue of $\mathbf{Q}$, then we have[7]

$$\|(\mathbf{Q} + \lambda\mathbf{I})^{-1}\mathbf{c}\| = \|\mathbf{x}\|$$

which implies,

$$\frac{1}{\underline{\psi}(\mathbf{Q}) + \lambda}\|\mathbf{c}\| \geq \|\mathbf{x}\| \tag{2.12}$$

[6] We use ψ to represent eigenvalues instead of the more familiar λ so that we can retain λ for Lagrangian multipliers.

[7] We require either $\lambda > 0$ or $\mathbf{Q}$ invertible for this step.

and thus,

$$\underline{\psi}(\mathbf{Q}) + \lambda \leq \frac{\|\mathbf{c}\|}{\|\mathbf{x}\|} \tag{2.13}$$

finally,

$$\lambda \leq \frac{\|\mathbf{c}\|}{\|\mathbf{x}\|} - \underline{\psi}(\mathbf{Q}) \tag{2.14}$$

If a non-zero λ^* satisfies the KKT conditions, then the corresponding $\mathbf{x}^*$ must have $\|\mathbf{x}^*\| = r$. If we do not have access to the eigenvalues of $\mathbf{Q}$, then we can drop the $\underline{\psi}(\mathbf{Q})$ term from (2.14) which relaxes the bound.

We now have a range of possible values for λ^* and, because strict complementarity holds for this problem, we know λ^* will be unique. We can test an arbitrary λ for optimality by solving (2.4a) for $\mathbf{x}$ and then checking its norm. If $\|\mathbf{x}\| = r$, then $\mathbf{x}$ is the optimal solution to the problem. This suggests a bisection algorithm for solving (2.9). Assume that we wish to find λ^* to within ϵ of the exact solution. The bisection algorithm, presented in [4], proceeds as follows:

1. set $\lambda_{\text{low}} = 0$ and $\lambda_{\text{high}} = \frac{\|\mathbf{c}\|}{r}$.

2. set $\lambda = \frac{1}{2}(\lambda_{\text{low}} + \lambda_{\text{high}})$.

3. if $(\lambda_{\text{high}} - \lambda_{\text{low}}) \leq \epsilon$ exit procedure, otherwise go to step 4.

4. solve $(\mathbf{Q} + \lambda\mathbf{I})\mathbf{x} = -\mathbf{c}$ for $\mathbf{x}$

5. if $\|\mathbf{x}\| > 1$, then set $\lambda_{\text{low}} = \lambda$, otherwise set $\lambda_{\text{high}} = \lambda$.

6. goto step 2

If, at termination, λ is within ϵ of zero, the algorithm has selected a point strictly inside the sphere constraint. Note that the solution to step 4 when $\lambda = 0$ is simply the stationary point of $q()$.

If a transformation was applied to obtain the formulation in (2.9), the solution to the original problem can be obtained by back-solving through the transformation once $\mathbf{x}$ is found.

This algorithm requires at least $\log_2 \frac{\|\mathbf{c}\|}{r} - \log_2 \epsilon$ iterations to converge to the desired precision. If the eigenvalues of $\mathbf{Q}$ are known, and $\mathbf{Q}$ is non-singular, this can be reduced. However, to eliminate one iteration, $\underline{\psi}(\mathbf{Q})$ must be greater than $\frac{\|\mathbf{c}\|}{2r}$ which is unlikely if $\|\mathbf{c}\|$ is large. Each iteration requires a single linear system solution. Because we assumed $q()$ was convex, $(\mathbf{Q}+\lambda\mathbf{I})$ must be symmetric positive semi-definite which implies fast Cholesky factorizations can be used to perform the matrix inversion required in step 4.

Alternatively, we can consider restricting the search to the surface of the sphere. Changing to an equality constraint changes the necessary conditions in (2.10a)-(2.10e) available to characterize a solution. Most significantly, the parameter λ^* is no longer restricted to be positive. A direct result of allowing λ^* to be negative is that (2.11) is no longer always positive semi-definite. In fact, for $\lambda < 0$, the second order term $(\mathbf{Q}+\lambda\mathbf{I})$ is negative definite with λ being the minimum eigenvalue. The second order condition must now be tested at a given point rather than assumed satisfied. More important than the necessary conditions, there is a fundamental change in the structure of the problem. An equality constraint specifies the surface of a sphere which is *not* a convex set. Thus, we can no longer use the observation that all local solutions of convex problems are also global solutions. To deal with these two changes we need methods to establish the new range for λ^*, and to determine optimality once it is found.

Fortunately we can educe from the structure of the problem some helpful facts about λ^*. First, we establish that for λ large enough, the norm of the solution to (2.10a) is less than r. We obtain this from (2.12) in the form of an upper bound. Specifically, by setting $\|\mathbf{x}\| = r$, we have

$$\frac{1}{\underline{\psi}(\mathbf{Q}) + \lambda} \|\mathbf{c}\| \leq r$$

so that,

$$\lambda \leq \frac{\|\mathbf{c}\|}{r} - \underline{\psi}(\mathbf{Q})$$

as long as $\lambda \geq -\underline{\psi}(\mathbf{Q})$.

Next, we want to establish a lower bound for λ^*. That is, we must establish that, for λ between some lower bound and the upper bound derived above, a solution to (2.10a) with norm r exists. To do this, we reconsider taking the norm of both sides of a rearranged (2.10a)

$$\|\mathbf{x}\| = \|(\mathbf{Q} + \lambda\mathbf{I})^{-1}\mathbf{c}\|$$

It is well known that the eigenvalues of $(\mathbf{Q} + \lambda\mathbf{I})^{-1}$ are $\frac{1}{\psi(\mathbf{Q})+\lambda}$ and that the eigenvectors are the same as those of $\mathbf{Q}$. Let $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ represent the eigenvectors of $\mathbf{Q}$ where y_i is matched to $\psi_i(\mathbf{Q})$. Note that, as λ approaches $-\psi_i(\mathbf{Q})$, the absolute value of $\frac{1}{\psi_i(\mathbf{Q})+\lambda}$ increases indefinitely for all $i \in \{1, \ldots, n\}$. Clearly then, the norm of the product $(\mathbf{Q} + \lambda\mathbf{I})^{-1}\mathbf{y}_i$ increases indefinitely as well. Because they must span $\mathbb{R}^n$, we can write $\mathbf{c}$ as a linear combination of the eigenvectors of $\mathbf{Q}$. Thus, as λ approaches the negative of any $\psi(\mathbf{Q})$ for which $\mathbf{c}^t\mathbf{y}_i \neq 0$, the norm of the solution vector $\mathbf{x}$ will increase indefinitely as well. Care must be taken to avoid situations in which the optimal solution is close to the eigenvalue. In these cases, the transformation is nearly singular. We can use a method such as `condest()` in Matlab (due to [7]) to detect these cases.

Clearly, the norm of $\mathbf{x}$ becomes larger than r when, from the set of eigenvectors of $\mathbf{Q}$ to which $\mathbf{c}$ is linearly dependent, λ approaches the the negative of the minimum corresponding eigenvalue. We use the non-standard notation $\psi^{\not\perp}(Q)$ to label this eigenvalue so that the importance of the linear dependence is emphasized. With assumed continuity, we know that somewhere in the range $-\underline{\psi}^{\not\perp}(\mathbf{Q}) \leq \lambda \leq \|\mathbf{c}\| - \underline{\psi}(\mathbf{Q})$ there must be a solution to (2.10a) whose norm is exactly r. Therefore we can still use the bisection algorithm to search for it. The problem then becomes whether the local solution found this way is a maximum or a minimum.

We must appeal to the second order condition which states, for $(\mathbf{x}^*, \lambda^*)$ a solution to the KKT conditions, $\mathbf{x}^*$ is a local minimum if

$$\mathbf{d}^t\mathbf{Q}\mathbf{d} + \lambda\|\mathbf{d}\|^2 \geq 0, \quad \forall\mathbf{d}, \mathbf{d}^t\mathbf{x}^* = 0$$

Because $\mathbf{Q}$ is positive semi-definite, the above condition will be satisfied whenever $\lambda \geq 0$. When λ is less than zero, we can use the *Rayleigh-Ritz* theorem which give us, for $\mathbf{Q}$ symmetric rational,

$$\mathbf{w}^t\mathbf{Q}\mathbf{w} \geq \underline{\psi}(\mathbf{Q})\|\mathbf{w}\|^2, \quad \mathbf{w} \neq 0 \tag{2.15}$$

and, for $\mathbf{y}_i$ the eigenvector which corresponds to the eigenvalue $\psi_i(\mathbf{Q})$,

$$\mathbf{w}^t\mathbf{Q}\mathbf{w} \geq \psi_i(\mathbf{Q})\|\mathbf{w}\|^2, \quad \mathbf{w} \neq 0, \mathbf{w} \perp \mathbf{y}_1, \ldots, \mathbf{y}_{i-1} \tag{2.16}$$

Thus, for $\lambda > -\underline{\psi}^{\not\perp}(\mathbf{Q})$ the second order condition is satisfied. Therefore, the local solution obtained by the bisection algorithm on the new range for λ^* will always be a local minimum. As is normal for non-convex problems, it may not be possible to establish a range for the global solution in this manner. Fortunately, we have a practical method which will always find a good solution in the sense that it minimizes the problem locally.

Chapter 3

Transmitter Based CDMA Optimization

In communications engineering, multiple access refers to techniques which allow a single channel to be shared amongst many users. Multiple access has great importance for commercial wireless communication because of limited spectrum allocation and the large dense populations of potential users in urban centers. Code-division multiple access (CDMA) is a powerful multiple access technique based on *spread-spectrum modulation* which underlies many third-generation cellular network standards. CDMA has many beneficial features for commercial cellular use including is its ability to reject unintentional interference from users sharing a channel, and multipath rejection. Borrowing from [8], spread-spectrum modulation can be defined in two parts as: a) a means of transmission in which more bandwidth is used than the minimum required to transmit the data, and b) a method which spreads the data through the spectrum by use of a code which is independent of the data stream. In CDMA systems, extra bandwidth is consumed to obtain the beneficial interference suppression side-effects. This complicates the relationship between bandwidth, power, and error performance resulting in an environment with many potential applications for quadratic optimization.

For cellular wireless systems, direct-sequence CDMA (DS-CDMA) is the most popular CDMA technique. In DS-CDMA, each user's data is multiplied by a distinct code waveform and broadcast simultaneously. A single user's data is recovered from the overlapped signals by correlating the received waveform with the desired user's code waveform. If the set of code waveforms are orthogonal, each recovered symbol is completely free of interference caused by channel sharing; however, it is not always possible or even desirable to have orthogonal code waveforms. Instead, the code waveforms are often cross-correlated which causes the recovered symbols to contain components of data from multiple users. This is called *multiple access interference* (MAI), and is the major limiting factor for DS-CDMA capacity.

DS-CDMA capacity can be increased by improved detection. It is well known that the conventional single-user DS-CDMA detector is sub-optimal because it treats the received signal as a single user broadcasting through Gaussian noise. In reality, DS-CDMA MAI is highly structured and therefore highly non-Gaussian. *Multiuser detection* (also known as joint detection or interference cancelation), exploits the structure of multiuser interference to improve DS-CDMA channel capacity at the cost of additional processing overhead.

Traditionally, this processing overhead is incurred at the receiver in the form of complex detectors. This is motivated by Verdú's early discovery of the optimal uncoded detector [9]. The optimal CDMA detector requires the full set of correlator outputs which are available only at the receiver. Because the optimal detector was found to have unacceptable complexity, most subsequent research has attempted to find efficient approximations of it. These approximations maintain the same structure as the optimal detector and therefore the same receiver dominant processing distribution.

Recently, approaches have been proposed which transfer some of the processing burden from the receiver to the transmitter. Although such methods cannot be optimal, they can be appropriate when heavy processing burdens at the receiver are unacceptable. The forward link (base-station to mobile handset) of a cellular telephone system is an excellent example. Methods such as those in [10], [11], and [12], allow some of the processing burden to be transferred from the resource-constrained handsets to the base-station where the cost of processing resources is lower.

Here we propose a new method, similar to one in [10], to achieve receiver to transmitter processing transfer which will be motivated by the goal of achieving the simplest possible receiver structure. The proposed technique is designed to be implemented with the techniques introduced in chapter 2. We begin with a brief review of DS-CDMA research and a description of the proposed method as it relates to previous research. This is followed by a system model in section (3.2) and the proposed method in (3.3). Numerical results follow in (3.4).

3.1 Review

Verdú introduced the optimal uncoded CDMA detector in his Ph.D thesis [9] and discussed it further in later papers [13, 14]. His derivation was based on a bank of matched filter receivers (one for each user) and maximum likelihood sequence detection of asynchronous signals in additive white Gaussian noise (AWGN). These works exposed a major limitation of multiuser detection; the optimal detector was shown to have a complexity which was exponential in the number of simultaneous users. This result directed research toward sub-optimal detectors with lower complexity.

One class of sub-optimal detectors are the linear detectors, so called because they apply a linear transformation to the detected decision statistics. In [15], optimal linear detection was discussed and again found to have exponential complexity. This necessitated sub-optimal linear detection, perhaps the the simplest form of which is the *decorrelating detector*. This detector completely removes multiuser interference by applying the inverted cross-correlation matrix to the set of correlator outputs. Although optimal in a noiseless channel, the decorrelating detector has the unfortunate side effect of *noise enhancement*. The transformation it applies increases the relative power of the received noise waveforms in the detector outputs[1]. In low signal to noise ratio (SNR) environments, the decorrelating detector, which is optimized to reject multiuser interference, performs *worse* than the conventional detector which is optimized to reject environmental noise.

Other linear detectors attempt to achieve a trade-off between the noise rejec-

[1] In the literature, this phenomenon is occasionally called signal attenuation instead.

tion of the conventional detector and the interference rejection of the decorrelating detector. One such detector is the linear minimum mean-square error (MMSE) detector [15]. Given knowledge of the received SNR level, this detector selects a vector of symbols with the minimum likelihood of error. The MMSE detector is efficient in general, and can retain noise rejection in low SNR environments while still achieving the interference rejection of the decorrelating detector when SNR is high.

Several works have proposed non-linear decision-driven detectors which use decisions on the bits of interfering users in the decision making process for other users. These *feedback detectors*, which use techniques such as multi-stage detection, decision-feedback detection, and successive or parallel interference cancelation, require reliable initial bits to perform training on the current channel parameters. They are best suited to high SNR channels where power imbalances cause MAI to dominate. Duel-Hallen *et al.* [16] discusses many of these receivers[2].

Typically, multiuser DS-CDMA research has focused on receiver optimization. In the cellular forward link, this places the bulk of the processing burden on a highly resource-constrained mobile device. If some of the burden could be transferred to the base-station, this would allow the mobile receiver to implement new features, reduce cost, or increase battery life. Centralized processing also permits easier system management and allows the cost of equipment to be amortized by the service provider. For these reasons, forward-link *processing transfer* is a desirable objective.

If the performance loss over other multi-user detection techniques is small, one

[2] For a more complete discussion of multiuser detection, see [17].

might wonder what the maximum possible processing transfer would be. Clearly, the appropriate measure is the amount of processing required at the receiver. Because DS-CDMA signals are overlapped in time, a correlator will always be required. This is the obvious minimum amount of receiver processing required. If anything more than this is needed, then *maximum* processing transfer is not achieved.

One option for forward link processing transfer is to optimize the spreading sequences. Conventional DS-CDMA uses pseudo-noise (PN) sequences which can have random cross-correlations from symbol to symbol. By making deterministic selections, this source of interference can be reduced. Families of codes with useful properties are often used, such as those suggested in [18]. Alternatively, sequences can be designed to have low cross-correlations on a per-symbol basis as in [19]. In both cases, the base station is the best candidate to make the selection in the forward link which achieves processing transfer as defined above. Deterministic methods to select spreading sequences can achieve significant performance gains but have the drawback of requiring a method to indicate the selected sequences to the receiver. This latter process will require extra work and thus does not achieve maximum processing transfer.

In [10], the spreading sequences are assumed to be fixed. Instead, a linear transformation is added at the transmitter which minimizes the means-square error between the vector of data symbols and the set of detector outputs. Referred to as *transmitter precoding*[3], this per-symbol transformation achieves decorrellation of the received waveforms without the noise enhancement of the standard receiver

[3] In [12] the term *decorrelating prefilter* was used.

based decorrelator because decorrellation is performed before noise is added to the signal. Once applied, the transmitter-generated precoding transformation designs a signal which can be efficiently detected by a standard correlating receiver. Thus, transmitter precoding achieves maximum processing transfer[4]. The case where the precoding solution was constrained to have constant average transmitted energy was also studied in [10]. *Constrained* transmitter precoding was found to have similar performance to the unconstrained case for uncoded transmission. Moreover, it was posited that constrained transmitter precoding was too complex to be practical given the marginal expected benefits.

Transmitter precoding can be thought of as a form of fast power control. It was noted in [20] that forward link power control cannot increase overall system capacity because the reverse link capacity is usually lower before optimization. However, in modern cellular systems with combined voice and data streams, it can be expected that channel usage will be asymmetric. That is, although voice connections will require symmetrical channel use, data will flow primarily in the forward direction. Aside from the benefit of simple receivers, transmitter precoding improves downstream data capacity.

In this work, we consider two constrained forms of transmitter precoding which are similar to [10]. Specifically, we consider a form of constrained transmitter precoding for synchronous DS-CDMA in which the transmitted energy is *capped* or *fixed* for each symbol period. The MAI will be minimized in an MMSE sense

[4] In [11], this work was extended to jointly optimized linear transformations at both the transmitter and each receiver. However, given the processing transfer requirements, it is not of interest here.

subject to these constraints, thus we refer to the techniques as the "optimizing precoder" with either an inequality or equality constraint. In addition, the solution algorithm presented will provide an adjustable trade-off between performance and computational efficiency.

The focus will be on coded transmission, where it will be demonstrated that constrained transmitter precoding of all three kinds (that in [10], and the two presented here) performs significantly better than in the uncoded case and in fact *outperforms* the transmitter precoding in [10]. Because the goal of this work is to have as simple a receiver structure as possible, the receiver will perform decoding independently from detection even though it is well known that combined detection and decoding is more effective.

3.2 System Model

The system model presented here matches the forward link of a symbol and chip synchronous DS-CDMA system. We assume perfect forward-link power control, so that data is transmitted with amplitude A_i to user i. The transmitted signal, $x(t)$, of a synchronous DS-CDMA system with K active users and symbol duration T_b is

$$x(t) = \sum_{n=1}^{K} A_n s_n(t) b_n, \qquad 0 \leq t \leq T_b \tag{3.1}$$

where b_n and $s_n(t)$ are the n^{th} user's modulated data symbol and signature waveform respectively. We assume that the data symbols are binary and equi-probable.

Additionally, we assume that the signature waveforms are linearly independent, zero outside the range $[0, T_b]$, and normalized to have unit energy so that

$$\int_0^{T_b} s_n^2(t)\, dt = 1, \qquad 1 \le n \le K \tag{3.2}$$

In the additive white Gaussian noise (AWGN) channel, the received waveform is

$$r(t) = x(t) + n(t), \qquad 0 \le t \le T_b \tag{3.3}$$

where n(t) is a Gaussian process with zero mean and two-sided power spectral density of $\sigma^2 = N_0/2$. The matched filter output for user n is the correlator output y_n where

$$y_n = \int_0^{T_b} r(t)s_n(t)dt, \qquad 1 \le n \le K \tag{3.4}$$

If we combine the matched filter outputs to form the vector $\mathbf{y} = [y_1, \cdots, y_K]^t$, then the output of a bank of matched filters can be described in matrix notation as

$$\mathbf{y} = \mathbf{RAb} + \mathbf{n} \tag{3.5}$$

where $\mathbf{b} = [b_1, \ldots, b_K]^t$ is the vector of modulated data symbols, $\mathbf{n} = [n_1, \cdots, n_K]^t$ is a Gaussian noise vector whose elements have zero mean and covariance matrix $\mathbf{R}[\sigma_1^2, \ldots, \sigma_K^2]$, $\mathbf{A} = \text{diag}\{A_1, \ldots, A_K\}$ is the set of amplitudes, and finally, $\mathbf{R}$ is the $K \times K$ cross-correlation matrix for the current set of signature waveforms. The entries of $\mathbf{R}$ are defined as

$$R_{i,j} = \int_0^{T_b} s_i(t) s_j(t) dt \tag{3.6}$$

The detection strategy which gives the minimum bit error rate selects the vector $\mathbf{b}$ with maximum-likelihood given the set of observations $\mathbf{y}$. If the off-diagonal elements of $\mathbf{R}$ are zero (implying the spreading codes are orthogonal), and the data symbols are equi-probable (as assumed), then it is well known that the optimal decision rule is $\text{sgn}(\mathbf{y})$. This is equivalent to a zero threshold decision rule.

For simplicity, we consider a binary phase-shift keyed (BPSK) system. Both data and signature codes use antipodal modulation. Thus, the signature waveforms are composed of square waveforms called "chips" which have uniform duration T_c, where T_c is typically much smaller than T_b. With this structure, the normalized signature waveforms can be represented as binary vectors in L-space where $L = \frac{T_b}{T_c}$. That is, the $s_n(t)$ can be represented as the binary code vectors $\mathbf{s}_n \in \{\frac{-1}{\sqrt{L}}, \frac{1}{\sqrt{L}}\}^L$ and the collection of codes as the matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_K \end{bmatrix} \tag{3.7}$$

Note that $\mathbf{R} = M^t M$ where $\mathbf{M}^t$ denotes the transpose of $\mathbf{M}$ and that $\mathbf{R}$ is symmetric and positive semi-definite.

In this model, the signature codes are selected randomly. As discussed in [21], the justification for this is two-fold. First, a system with random codes will closely match the performance of a system that uses long pseudo-noise sequences to generate its signature codes (e.g.: IS-95 [22]). Second, a randomly coded system provides a lower bound for the performance that can be achieved. That is, random codes can be seen as averaging the performance of all deterministic code sequences.

3.3 Proposed Method

Our stated goal is to improve BER in a DS-CDMA channel with a conventional receiver by preprocessing the transmitted signal. There is an unlimited number of preprocessing operations that could be chosen. Fortunately, we need not select blindly. A common approach to estimate a data vector $\mathbf{b}$ based on the set of observations $\mathbf{y}$ is to choose the function $\hat{\mathbf{b}}()$ which minimizes the mean-square error

$$E[(\mathbf{b} - \hat{\mathbf{b}}(\mathbf{y}))^2] \tag{3.8}$$

Such a function achieves *maximum likelihood* decoding of $\mathbf{b}$ which minimizes the bit error rate. Thus, minimum mean-square error can be used as a criterion in the design of a preprocessor systems which attempt to reduce BER.

One option for preprocessing is transmitter precoding which applies a linear transformation $\mathbf{T}$ to the transmitted vector so that

$$\mathbf{y} = \mathbf{RTAb} + \mathbf{n} \tag{3.9}$$

It was shown in [10] that when the MMSE criterion is used to solve for $\mathbf{T}$ and expectation is taken with respect to $\mathbf{b}$ and $\mathbf{n}$ the result is $\mathbf{T} = \mathbf{R}^{-1}$.

Similarly, we can consider an arbitrary transformation on $\mathbf{b}$ which gives the real-valued vector $\mathbf{b}'$ so that

$$\mathbf{y}' = \mathbf{Rb}' + \mathbf{n} \tag{3.10}$$

where we have incorporated $\mathbf{A}$ into $\mathbf{b}'$. Using the MMSE criterion to select $\mathbf{b}'$, we obtain mean-square error ρ where

$$\begin{aligned} \rho &= E[\|\mathbf{Ab} - \mathbf{y}'\|^2] && (3.11) \\ &= E[\|\mathbf{Ab} - (\mathbf{Rb}' + \mathbf{n})\|^2] && (3.12) \end{aligned}$$

which, after taking expectation with respect to $\mathbf{n}$, gives

$$\rho = \|\mathbf{Ab} - \mathbf{Rb}'\|^2 \tag{3.13}$$

so that the optimization problem can be expressed as

$$\underset{\mathbf{b}'}{\operatorname{argmin}} \|\mathbf{Ab} - \mathbf{Rb}'\|^2 \tag{3.14}$$

where the solution $\mathbf{b}' = \mathbf{R}^{-1}\mathbf{Ab}$ is the same optimal solution as for transmitter precoding. If we assume $\mathbf{R}$ is non-singular[5], then (3.14) always has a solution.

Like the decorrelating detector, selecting $\mathbf{b}'$ or using the optimal $\mathbf{T}$ completely eliminates MAI. However, because the decorrelating operation occurs before transmission, this method actually *outperforms* the decorrelating detector. To see this, consider the output of the conventional detector for both situations

$$\begin{aligned} \mathbf{y}_{\text{dec}} = & \quad \mathbf{R}^{-1}\mathbf{RAb} + \mathbf{Rn} & = \mathbf{Ab} + \mathbf{Rn} \\ \mathbf{y}_{\text{pre}} = & \quad \mathbf{Rb}' + \mathbf{n} & = \mathbf{Ab} + \mathbf{n} \end{aligned} \tag{3.15}$$

The vector $\mathbf{y}_{\text{dec}}$ is the result of using the decorrelator and $\mathbf{y}_{\text{pre}}$ is the precoding result when the $\mathbf{b}'$ formulation is used. Because $\mathbf{R}$ is positive definite, the $\mathbf{Rn}$ term

[5] For the remainder we assume $\mathbf{R}$ is non-singular; however, this assumption can be replaced with the more general constraint that the k^{th} user signature code is not spanned by the other codes. The reader is referred to chapter 5 in [17] for methods to deal with this more general situation.

in $\mathbf{y}_{\text{dec}}$ causes noise enhancement. That is, the relative strength of the signal in $\mathbf{y}_{\text{dec}}$ is lower. Note that while precoding removes all MAI, it does not suffer from noise enhancement.

A side effect of precoding is fluctuating transmitted power. This can be handled, as in [10], by scaling the precoding result for each symbol period. It is natural, however, to consider imposing an energy constraint on (3.14). If we separate $\mathbf{R}$ into the encoding matrix $\mathbf{M}$ and correlating matrix $\mathbf{M}^t$, we can reformulate the optimization as

$$\begin{aligned} \text{minimize:} \quad & \|\mathbf{Ab} - \mathbf{M}^t\mathbf{Mb}'\|^2 \\ \text{subject to:} \quad & \|\mathbf{Mb}'\|^2 \leq r \end{aligned} \tag{3.16}$$

I refer to this method as the *optimizing precoder*. From the set of $\mathbf{b}'$ with transmitted power less than or equal to r, it selects the element which results in the minimum detector output distance due to MAI. The parameter r is held constant from symbol to symbol so instantaneous transmitter power is capped by a fixed value. By limiting transmitter power we constrain the system to include MAI in situations where eliminating it would require more power. Clearly, when SNR is high, this can only worsen performance as MAI will dominate BER. An alternative is to *fix* transmitted power which imposes the constraint $\|\mathbf{Mb}'\|^2 = \sqrt{r}$ on (3.16). Fixed transmitted power has practical benefits, but the solution to such a problem will always contain residual MAI.

There are some critical differences between (3.16) and the constrained optimization presented in [10], the most significant of which is the scope of the precoder design. In (3.16), a new transformation is generated for each symbol period, with expectation of MSE taken with respect to environmental noise only. In [10], expectation of MSE is taken with respect to **b** as well as **n** so that the current data vector is not a distinct influence on the design of the precoding transformation. Including **b** in the design of the transmitted vector cannot impair performance. Rather, we expect that including the state of the data vector will improve the result. Furthermore, (3.16) caps instantaneous power at each interval while in [10] *average* power is constrained. The latter method has more flexibility, but allows for undesirable spikes in power. Finally, (3.16) is an optimization over a vector rather than a matrix. This brings the problem into the realm of numerical optimization so that techniques like those presented in chapter 2 can be used.

The constrained (or optimizing) precoder can be expected to perform worse than unconstrained precoding whenever SNR is high in an uncoded system. The situation changes, however, when forward error correction (FEC) is used. FEC coders achieve a form of multi-user detection by spreading bit energy over several symbol periods and treating intervals with high instantaneous MAI as less reliable. This reduces the effect of MAI in the high SNR regions where MAI dominated uncoded performance. Transmitter precoding can be thought of as a form of fast power control which reduces the relative power assigned to highly interfering users. When FEC is used, highly interfering users become less important to total BER, but the unconstrained precoding algorithm will still attenuate those users the same amount. By including a power constraint, the maximum relative change in user power is restricted so that interference reduction is subordinated to power control which is appropriate for a coded situation. This conclusion will be borne out by simulation results.

Simulation will also show that the method specified by (3.16) performs better than the constrained method in [10]. The method specified by (3.16) permits only those solutions with low transmitted energy. In contrast, the constrained method in [10] constrains average energy which allows instantaneous energy to exceed the constraint. However, when (3.16) uses an equality constraint to specify fixed transmitted power, performance drops, showing that MAI has still has a prominent role.

Numerical optimization techniques are required to solve (3.16). The objective function is squared distance due to MAI while the constraint function specifies an ellipsoid whose surface is the set of signal vectors which have a transmitted power level of r. Both functions are clearly convex, thus (2.9) is an appropriate optimization structure. We put (3.16) into the form of (2.9) by first introducing vector $\mathbf{x}$ where

$$\mathbf{x} = \frac{1}{\sqrt{r}}\mathbf{M}\mathbf{b}' \tag{3.17}$$

After dropping the exponent on the constraint, substituting $\mathbf{x}$ into (3.16) gives us

$$\begin{aligned} \text{minimize:} \quad & \|\mathbf{A}\mathbf{b} - \sqrt{r}\mathbf{M}^t\mathbf{x}\|^2 \\ \text{subject to:} \quad & \|\mathbf{x}\| \leq 1 \end{aligned} \tag{3.18}$$

Expanding the new objective function results in

$$\mathbf{b}^t\mathbf{A}^t\mathbf{A}\mathbf{b} - \sqrt{r}\mathbf{b}^t\mathbf{A}^t\mathbf{M}^t\mathbf{x} - \sqrt{r}\mathbf{x}^t\mathbf{M}\mathbf{A}\mathbf{b} + r\mathbf{x}\mathbf{M}\mathbf{M}^t\mathbf{x} \tag{3.19}$$

To put this in the form of (2.9), let $\mathbf{Q} = 2r\mathbf{M}\mathbf{M}^t$, $\mathbf{c} = -2\sqrt{r}\mathbf{M}\mathbf{A}\mathbf{b}$, and delete the constant term. At this point, the solution algorithm can be applied directly to obtain $\mathbf{x}^*$. The final result $\mathbf{b}^*$ is retrieved by back-solving the linear system in (3.17) or solving $\mathbf{M}^t\mathbf{M}\mathbf{A}\mathbf{b}^* = \sqrt{r}\mathbf{M}^t\mathbf{x}$. The latter can be rewritten as $\mathbf{R}\mathbf{A}\mathbf{b}^* = \sqrt{r}\mathbf{M}^t\mathbf{x}$ which always has a solution when $\mathbf{R}$ is invertible.

The complexity of the solution algorithm depends on the precision parameter (which affects the number of iterations), and the dimension of $\mathbf{Q}$. Because $\mathbf{Q}$ is $L \times L$, the proposed solution has complexity which is *constant* in the number of users. In practical systems, the chip rate is typically fixed. This is an important performance result for practical systems which are the target of this method.

The Hessian $\mathbf{Q}$ has an unusual structure. Like $\mathbf{R}$, it contains cross-correlations, in this case between the signature code bits at each chip interval. Thus, like $\mathbf{R}$, $\mathbf{Q}$ is symmetric positive semi-definite. However, whenever the number of active users exceeds L, matrix $\mathbf{Q}$ fails to be invertible. This causes the solution algorithm to fail if λ equals zero at any step. One possible solution is to generate linearly independent signature codes for inactive users. Alternatively, the solution algorithm can be tweaked to avoid testing $\lambda = 0$.

3.4 Numerical Results

In this section, we analyze the performance of both the equality and inequality-constrained versions of the optimizing precoder in a simulated DS-CDMA environment. Simulation results will be compared with those from conventional systems as well as with both the constrained and unconstrained transmitter precoding methods of [10]. It will be shown that, in environments utilizing error-correcting codes, energy-constrained methods out-perform all other methods considered here. Furthermore, it will be shown that the energy-constrained methods proposed in this thesis out-perform the energy-constrained method from [10].

3.4.1 Simulation Setup

The channel model used for simulation matches the forward link of a single-cell synchronous DS-CDMA system using BPSK modulation as specified in section 3.2. The system chip rate is set to 32 in all cases which provides a reasonable approximation of practical systems (see [22]). Both coded and uncoded systems are considered with coded systems using the rate 1/3 convolutional code from the IS-95 cellular communication standard (see [22]). Receivers use a standard Viterbi decoder which is supplied with soft outputs from the detector. Only active users transmit and there is no provision for ARQ packet re-transmission schemes.

The simulator generates independent pseudo-random data streams for each user which are independently coded if ECC is used. Another independent pseudo-random bit stream, shared by all users, is used to generate the spreading codes. After modulation and spreading, AWGN is added to the combined signals. A bank of independent detectors then produces soft-outputs from the received signals which are either fed to a Viterbi decoder or used to make hard decisions for the original data bits.

Two types of power-control are investigated: an idealized environment with equal power users, and a more realistic near/far environment with uniformly distributed users and perfect power-control. The simulated near/far environment uniformly positions each user on the range $(1, 4)$ so that the maximum power ratio $(A_i/A_j)^2$ is always 16. For example, four active users would be positioned at the distances 1, 2, 3, and 4 and would have the same assigned amplitudes (A_i). This is a reasonable approximation of a practical system in which users are roaming in a hexagonal cell.

3.4.2 Optimizer Setup

With the structure of the simulation system in place, the optimal value of the power constraint r for the proposed method can be determined. With no closed form solution for the optimal r in (3.16), we must use simulation to discover it. By holding the system parameters and SNR constant while varying r, we hope to find a value which gives the minimum BER for each system configuration. Fortunately, BER will be smooth and continuous with respect to changes in r which will aid in the search.

To see this, consider an arbitrary solution to (3.16). Because transmitted energy is a quadratic function and therefore itself smooth and continuous, a small change in r requires only a small change in the optimal solution $\mathbf{b}'$ to achieve the desired transmitted energy. In turn, because MAI is also quadratic and therefore smooth and continuous, a small change in $\mathbf{b}'$ results in only a small change in MAI. As the change in r disappears, so will the change in the resulting MAI, which implies MAI is smooth and continuous with respect to r. Finally, because residual MAI dominates BER performance when SNR is held constant, we can expect BER to be smooth and continuous with respect to r as well.

This intuitive justification is supported by the simulation results shown in Figures 3.1 and 3.2. The results show the effect of changing r on the equality constrained optimizing precoder's performance when the system configuration and SNR level are held constant. In all eight curves, performance reaches a single global optimum with respect to r. This is a result of the optimization step which results in a different set of optimizer solutions for each r. One such set will have the minimum total MAI, implying that r must have a minimum value.

The optimum value of r was consistently lower for coded transmission than for uncoded. Moreover, the optimal r was much less affected by the number of users and changing channel conditions. Likely, this is due to the ability of FEC to spread bit energy over time. By averaging over several symbol periods, the effect of high instantaneous MAI is reduced and the benefit of tightly constrained transmitted power levels is enhanced.

Despite the effects of coding, the optimum r was not constant for different system configurations or even for different SNR levels. The system is required to adjust r constantly for changing conditions to achieve optimum performance. However, the set of optimum values for r need only be calculated once and stored

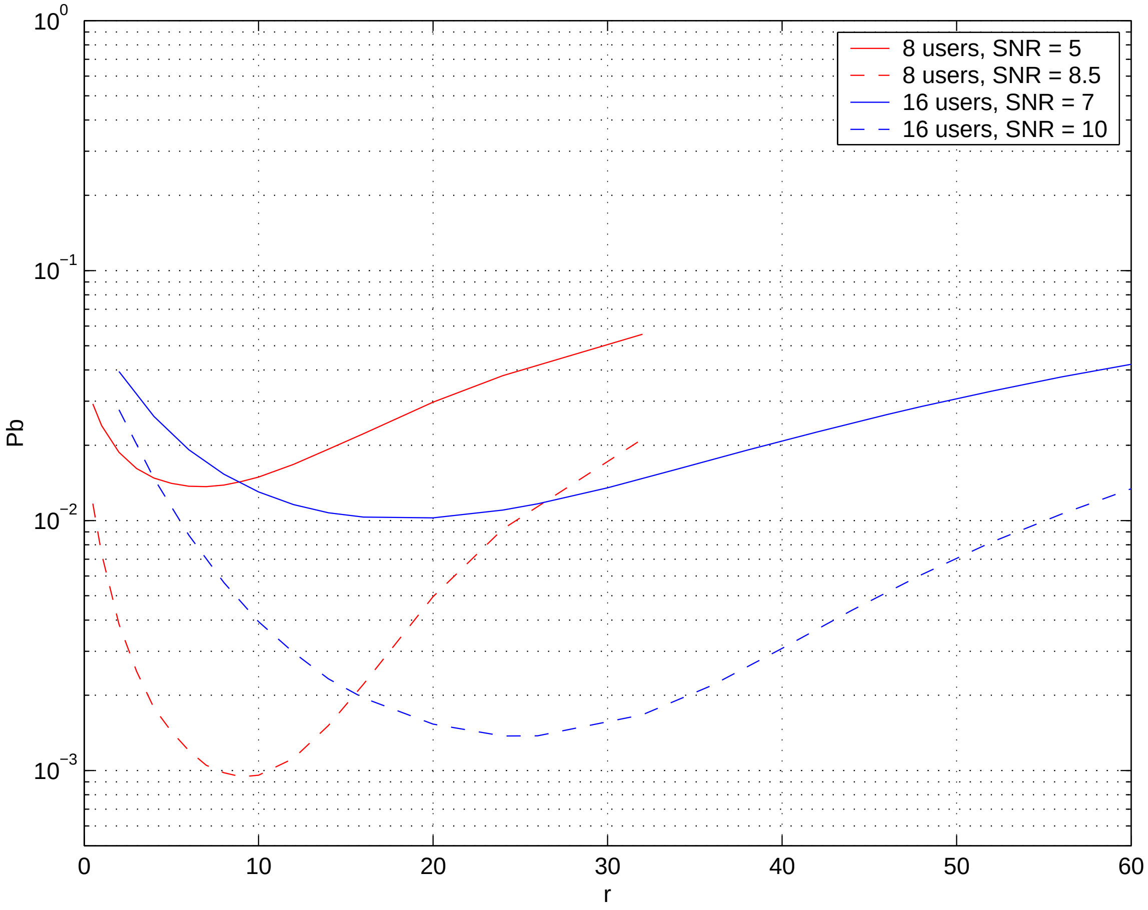

Figure 3.1: Uncoded bit error rate versus r for fixed user count and SNR

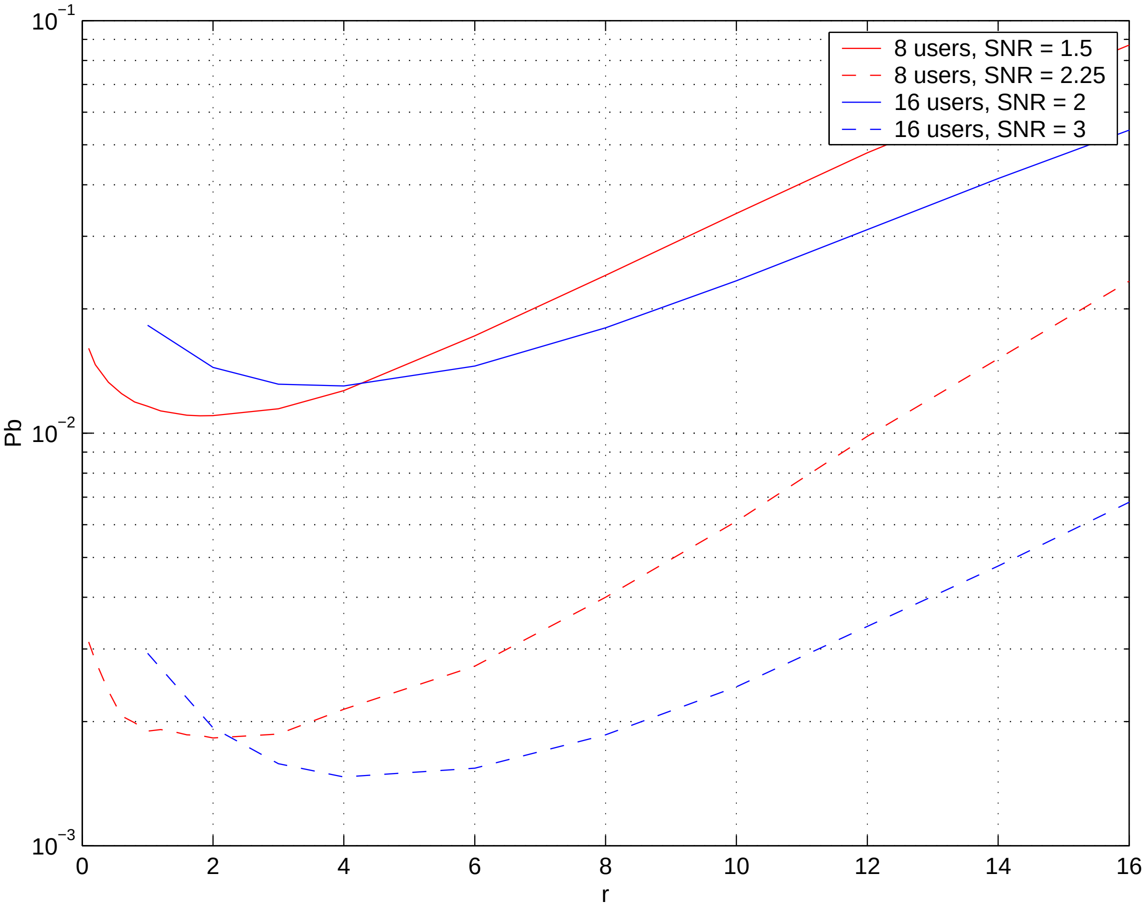

Figure 3.2: Coded bit error rate versus r for fixed user count and SNR

in a look-up table. Alternatively, r can be chosen to optimize performance at a certain SNR with only a small loss at other SNR levels due to the smoothness of the BER curve. Note that a conventional detector does not require knowledge of r to produce soft output metrics for this type of signal so that the modulator is free to adjust r as necessary.

3.4.3 Channel Performance Comparison

Results for several different DS-CDMA systems are shown below. Each simulated "system" consists of a modulator/detector pair and produces a single curve on each graph. Two conventional systems are considered: a system with a standard modulator and detector, and a system with a decorrelating detector. The former is labeled "Conventional System" and provides a baseline performance curve with no performance enhancing techniques. The latter is labeled the "Decorrelating Detector" and gives an example of a standard receiver-based performance enhancement technique. The two methods from [10] are labeled "Transmitter Precoding" and "Constrained Transmitter Precoding". They consist of the appropriate modulator with a standard detector. The optimizing methods proposed in this paper are labeled "Optimizing Precoder: $\leq$" and "Optimizing Precoder: $=$" for the inequality and equality-constrained problems respectively. Both are paired with a standard detector. When possible the single user case is also given for comparison.

In all cases, results for the proposed optimizing precoders are generated with a single value for r which is chosen to optimize performance at a BER of 10^{-3}. Precision (the parameter ϵ) for the optimizing modulators is set to 10^{-3} which corresponds to 10 iterations of the main loop of the solution algorithm.

Results for the various systems are grouped into “configurations” with a common active user count, a common power-control model, and the same presence or lack of ECC. In this manner, the relative strengths of the various methods are clearly displayed. Results are stated as the bit error rate averaged over all active users.

In Figures 3.3, 3.4, and 3.5, data is being transmitted uncoded with equal power to various numbers of users. In all cases, transmitter precoding gives the best raw performance at practical system error rates of 10^{-3} and below. As stated in [10], constrained transmitter precoding performs better for low SNR, but eventually crosses the unconstrained method as SNR rises. The optimizing precoders perform worse than constrained precoding for 8 users, but, as the numbers of users rises to 24, eventually overtake constrained precoding. However, all three energy-constrained methods eventually stop responding to rising SNR as residual MAI limits their performance.

In Figures 3.6, 3.7, and 3.8, the results are quite different. When channel coding is applied, the energy-constrained methods significantly out-perform unconstrained methods. With 24 users, constrained precoding gives a gain of approximately 3.5dB over unconstrained precoding. The proposed methods perform even better, achieving gains of 0.25dB, 0.5dB, and 0.75dB over constrained precoding with 8, 16, and 24 users respectively. The two optimizing methods themselves perform roughly identically in each case.

Figures 3.9 and 3.10 show the 8 and 16 user near/far cases respectively. Here the constrained methods perform closer to the unconstrained methods. A gain

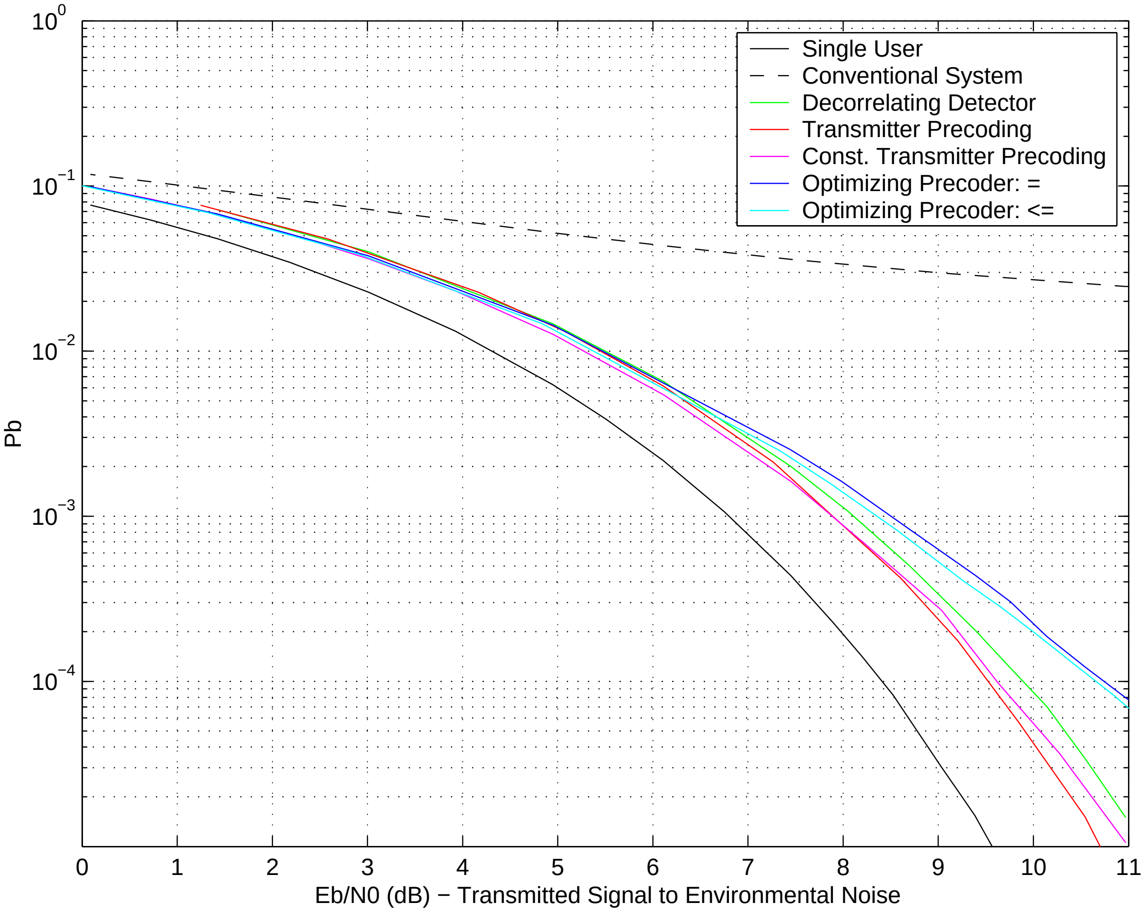

Figure 3.3: 8 Equal Power Users, Not Coded

is still achieved however, with the optimizing precoders only slightly better than constrained precoding and about 0.5dB better than unconstrained transmitter precoding at a BER of 10^{-3} and with 24 users.

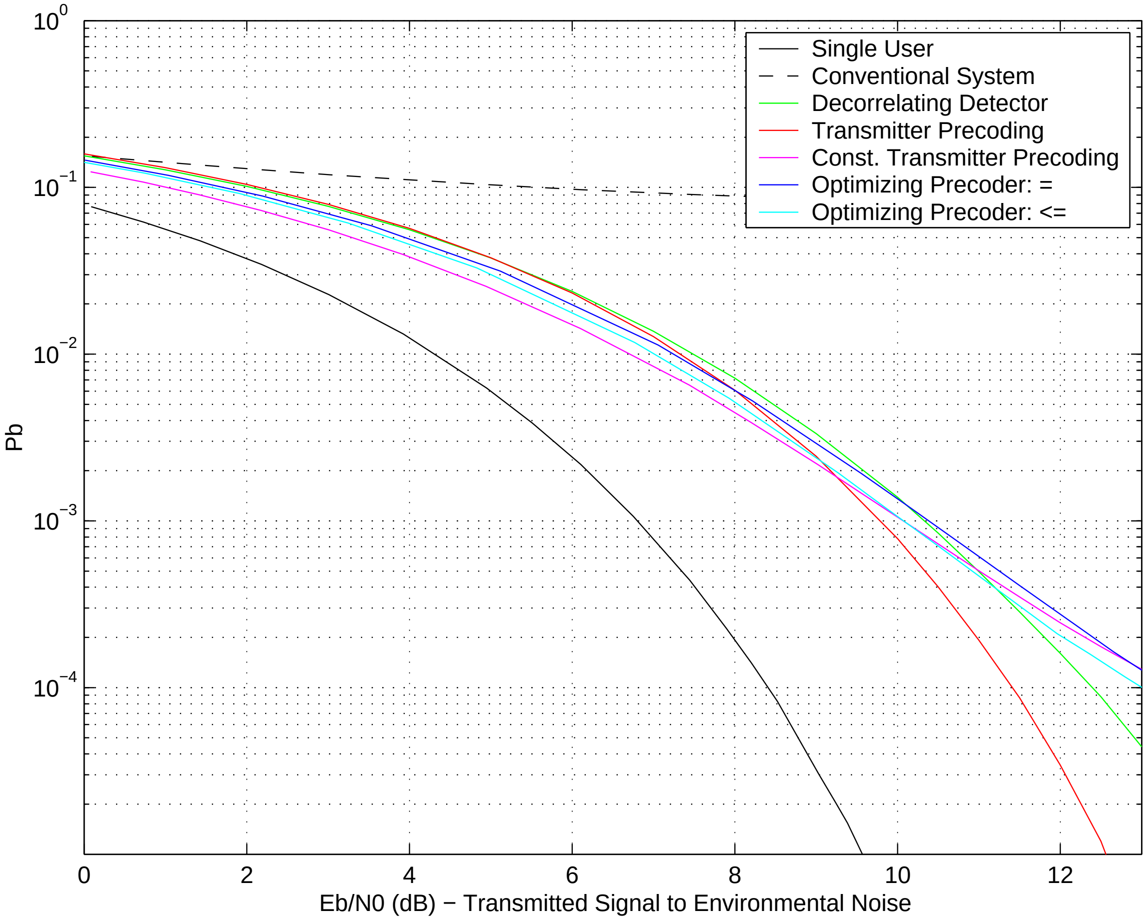

Figure 3.4: 16 Equal Power Users, Not Coded

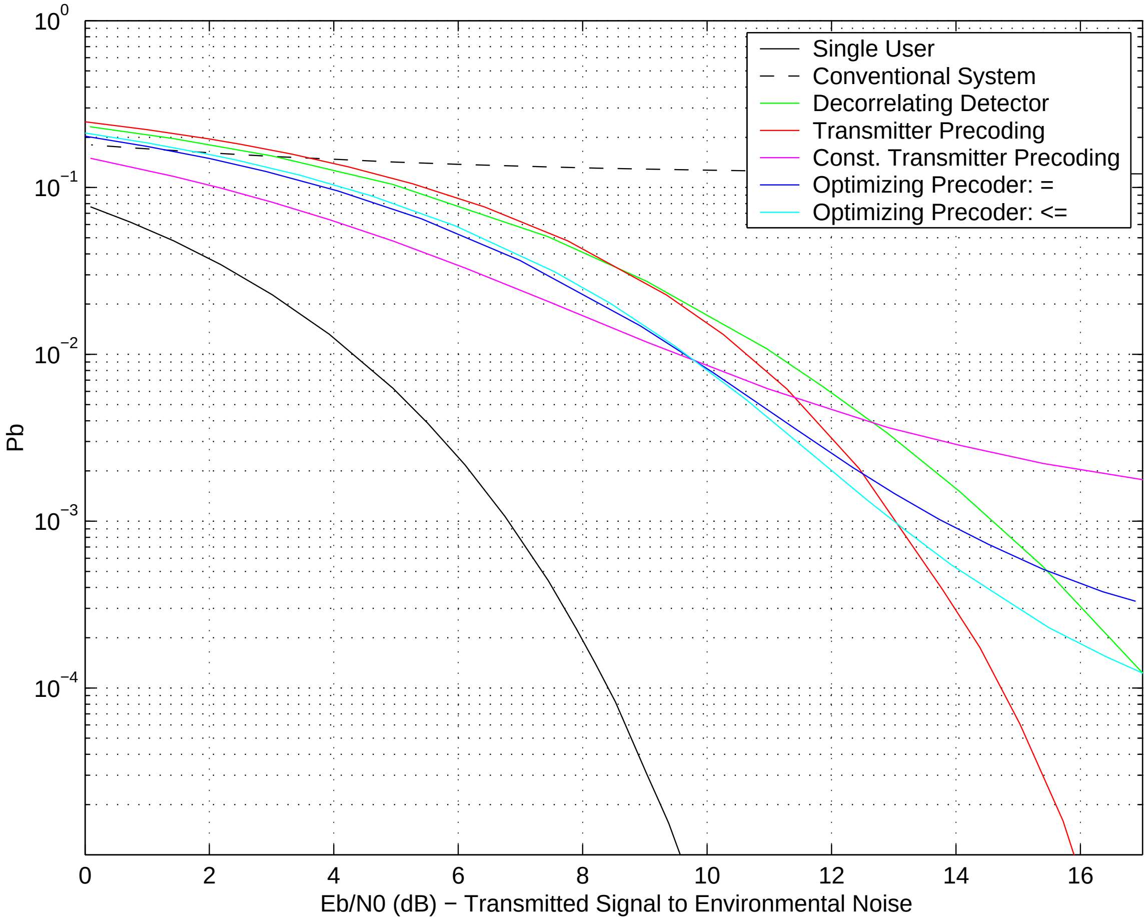

Figure 3.5: 24 Equal Power Users, Not Coded

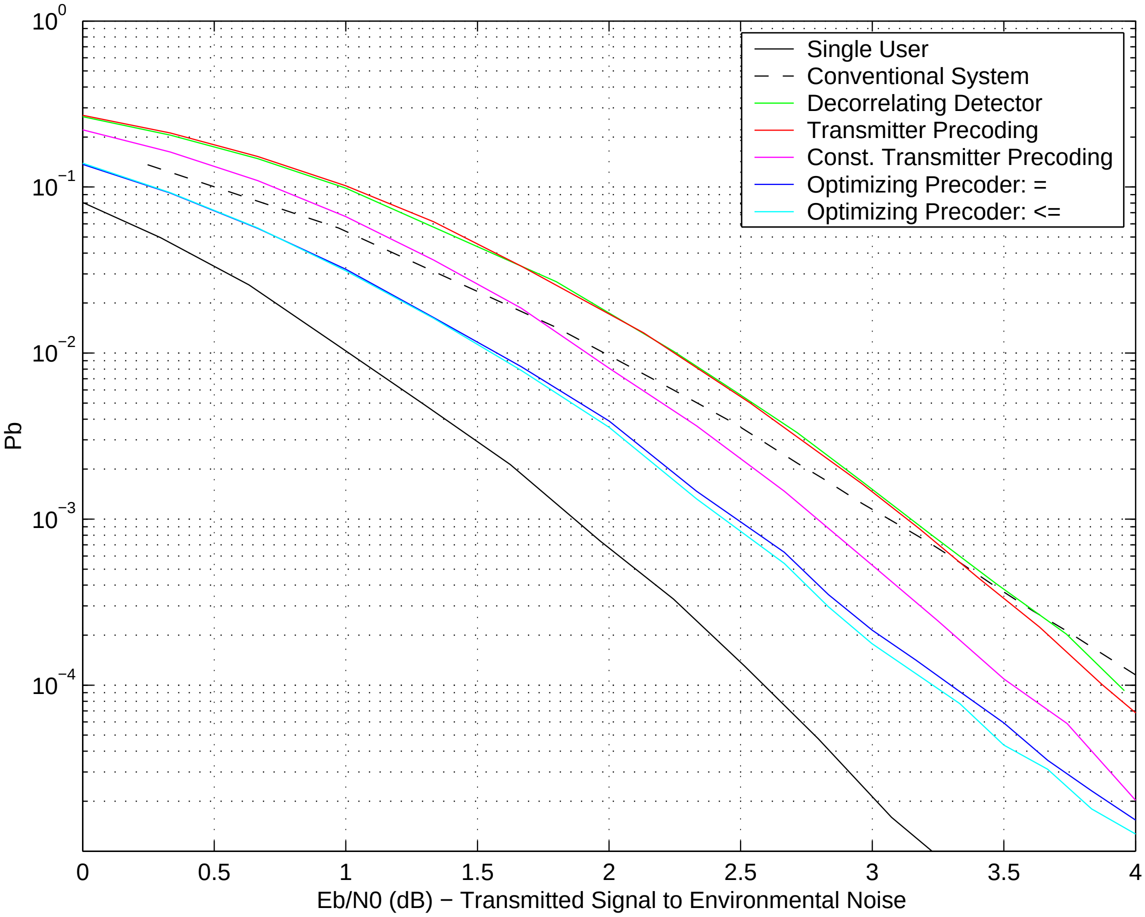

Figure 3.6: 8 Equal Power Users, Coded

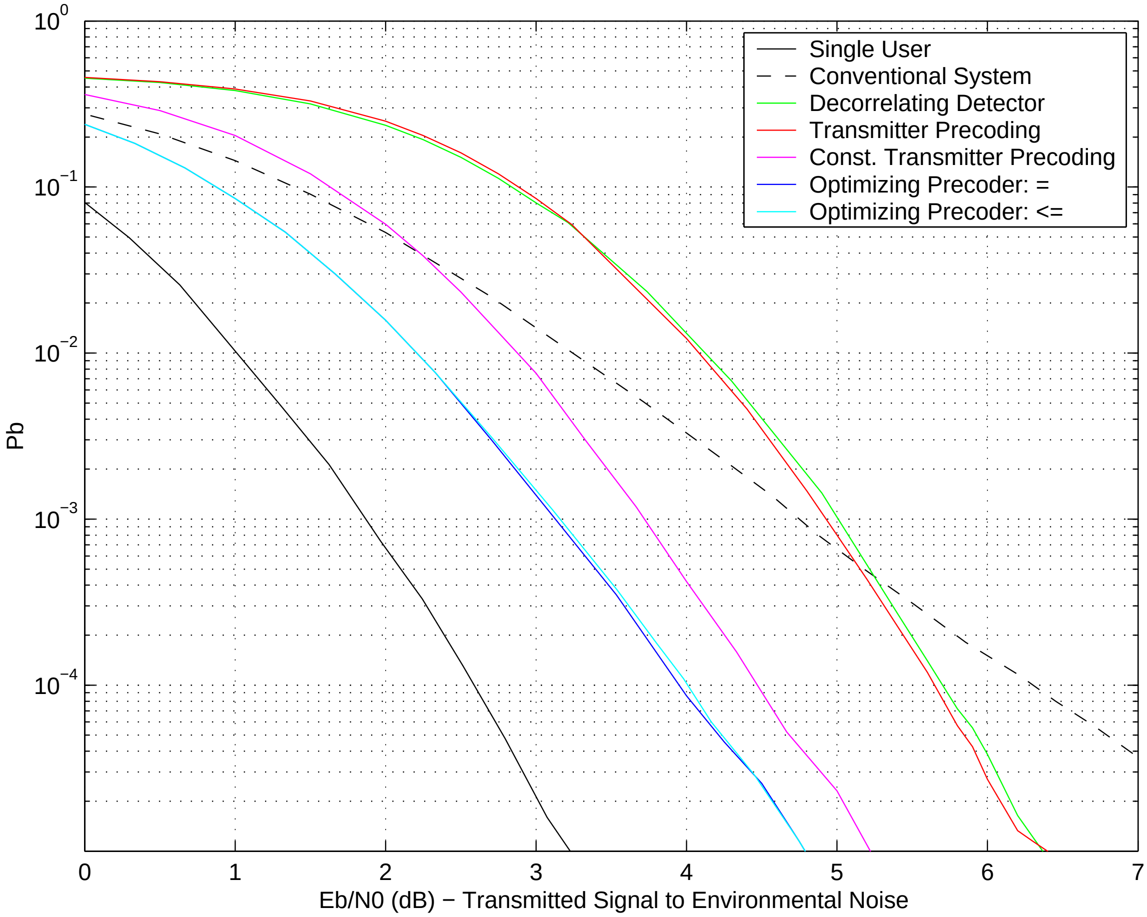

Figure 3.7: 16 Equal Power Users, Coded

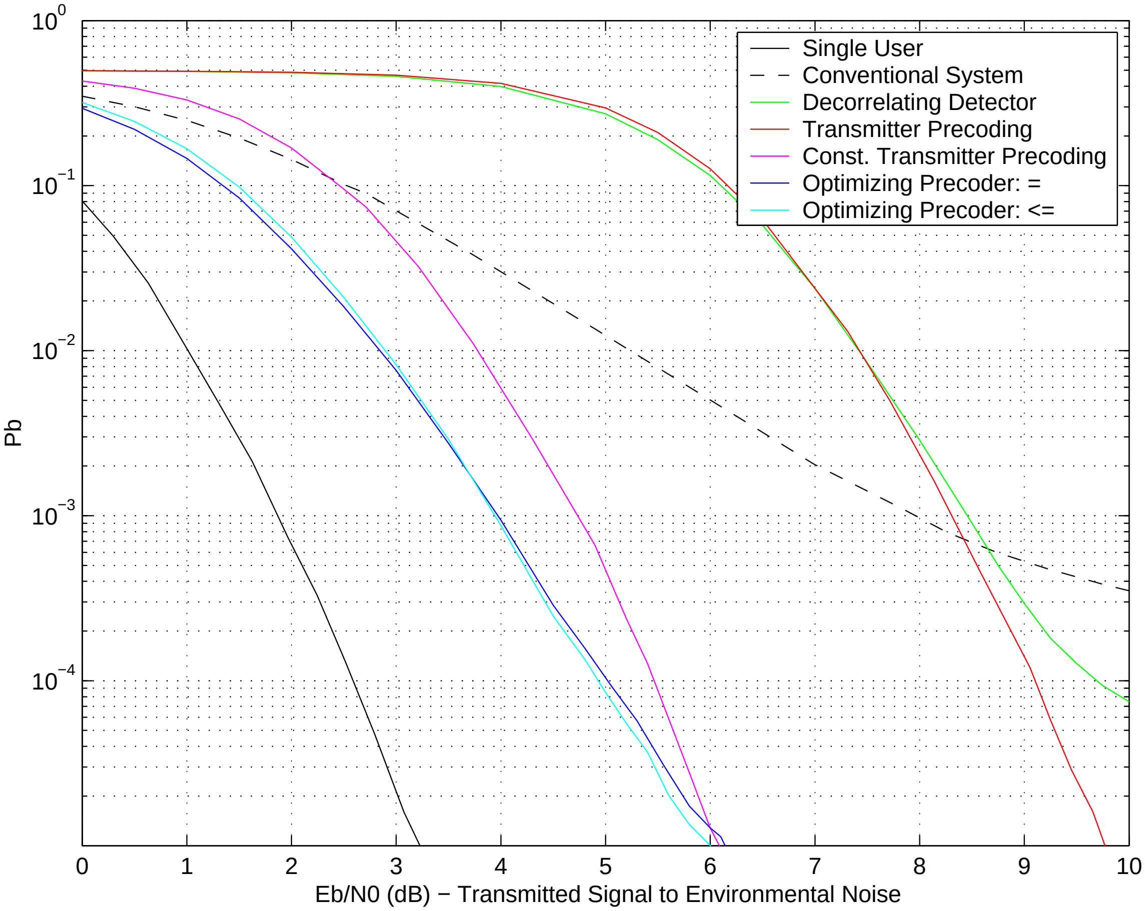

Figure 3.8: 24 Equal Power Users, Coded

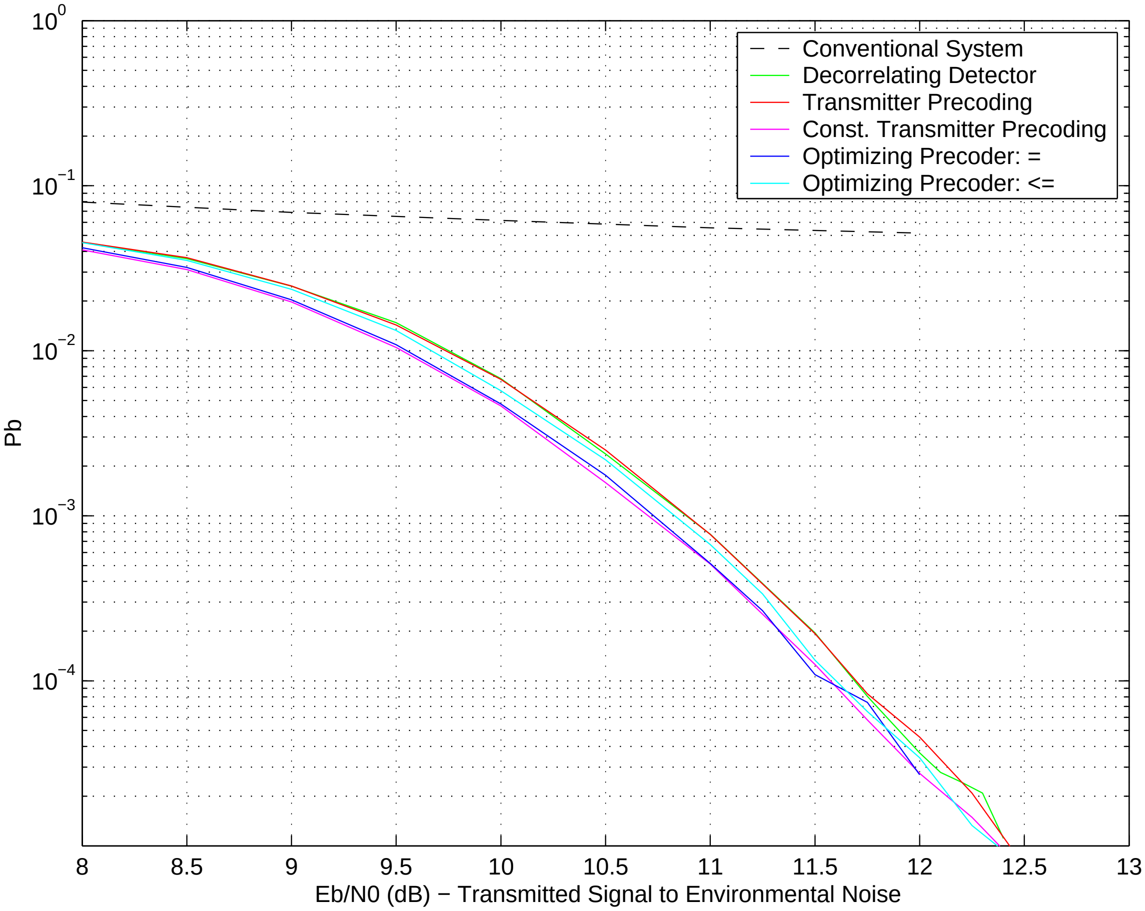

Figure 3.9: 8 Near/Far Users, Coded

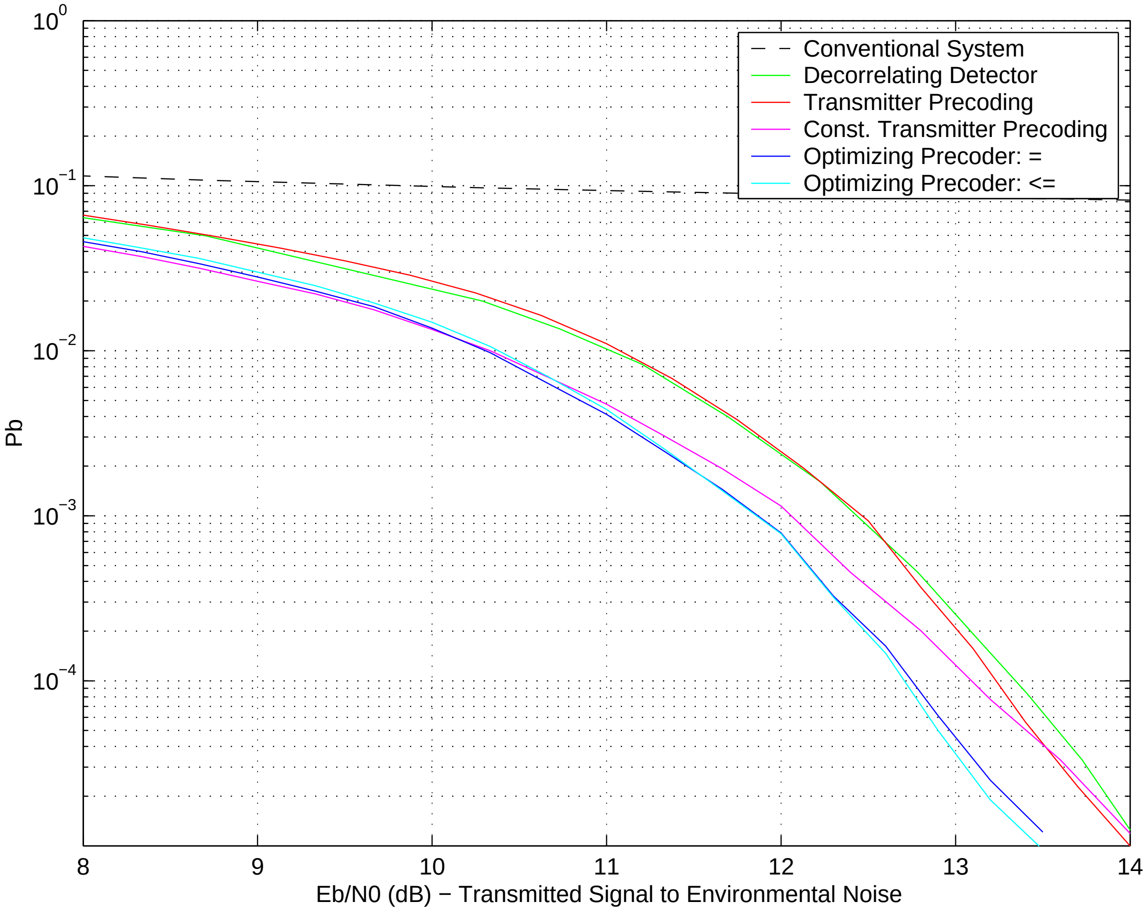

Figure 3.10: 16 Near/Far Users, Coded

3.4.4 Complexity Analysis

The computational complexity of the solution algorithm for both proposed methods is dominated by the desired accuracy of the solution and the cost of the matrix inversion in the body of the algorithm's main loop. However, only desired accuracy can be used to tune performance.

Accuracy directly translates into iterations of the main loop. Each single iteration divides the search range in half so that an accuracy of $1/2^n$ requires n iterations. In the following graphs, the performance curve for the inequality constrained optimizing precoder is regenerated with the number of iterations fixed at various levels[6]. Figure 3.11 shows that with 8 users and no coding the first iteration obtains all of the gain for the method. Figure 3.12 shows that for 16 uncoded users, the first iteration actually *outperforms* later iterations with early iterations performing better than more accurate and thus more costly solutions. We might have expected that more iterations would steadily improve performance because the accuracy of the solutions would improve. However, in uncoded situations the energy constraint is actually a hindrance so a less accurate solution can perform better.

More important for the goals of this thesis is coded performance, shown for the inequality constrained optimizing precoder in Figures 3.13 and 3.14 with 8 and 16 users respectively. From these graphs it is clear that that near-optimal performance is achieved on the second iteration, which means only two matrix inversions are required. This result dramatically improves the feasibility of these methods for

[6] The parameter r was tuned for optimal performance at a BER of 10^{-3} for each case

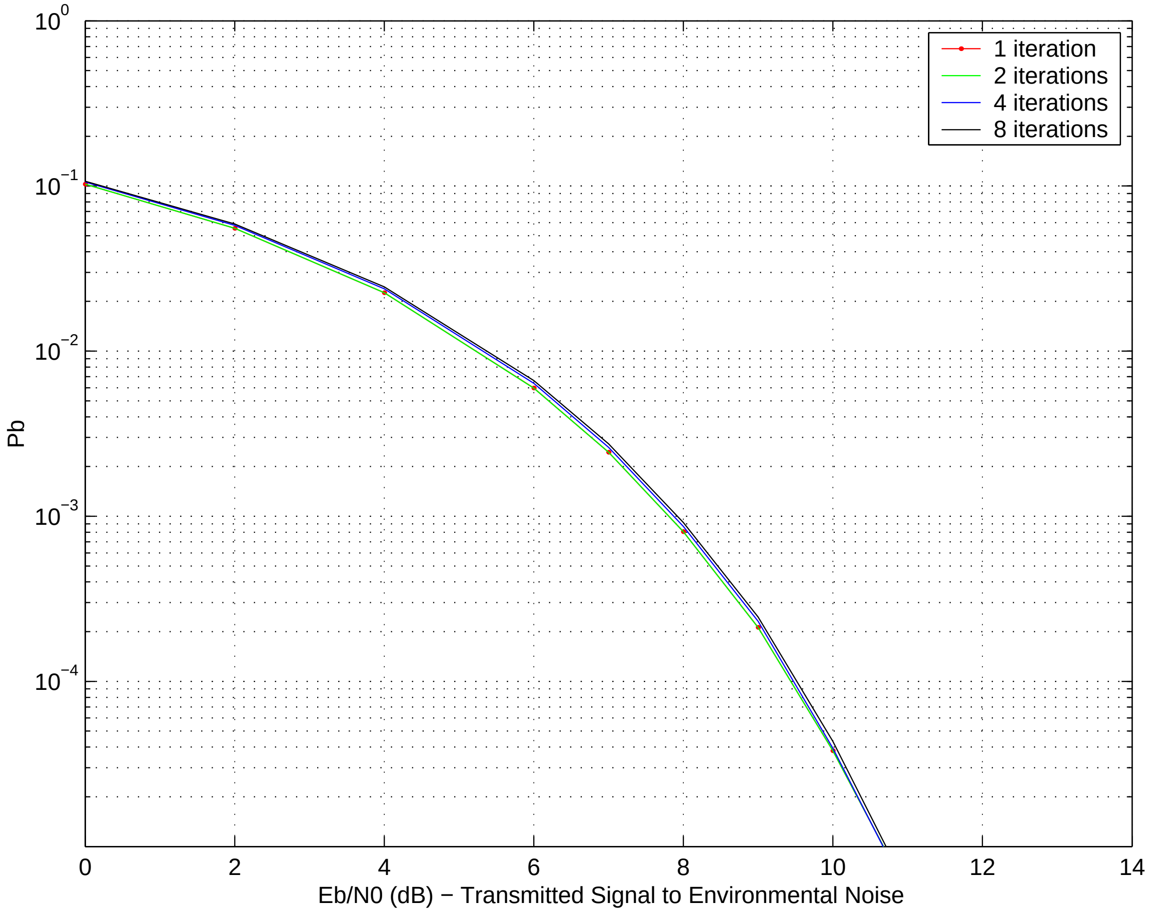

Figure 3.11: Complexity: 8 Equal Power Users, Not Coded

practical systems.

In all cases from the previous section, the inequality constrained optimizing precoder performed better than the equality constrained version. Furthermore, the inequality constrained method is less complex. Not only is it more accurate (because the search range is smaller), but it also only requires fast Cholesky factorizations whereas the equality constrained methods will occasionally require a slower inversion method when the matrix becomes negative definite.

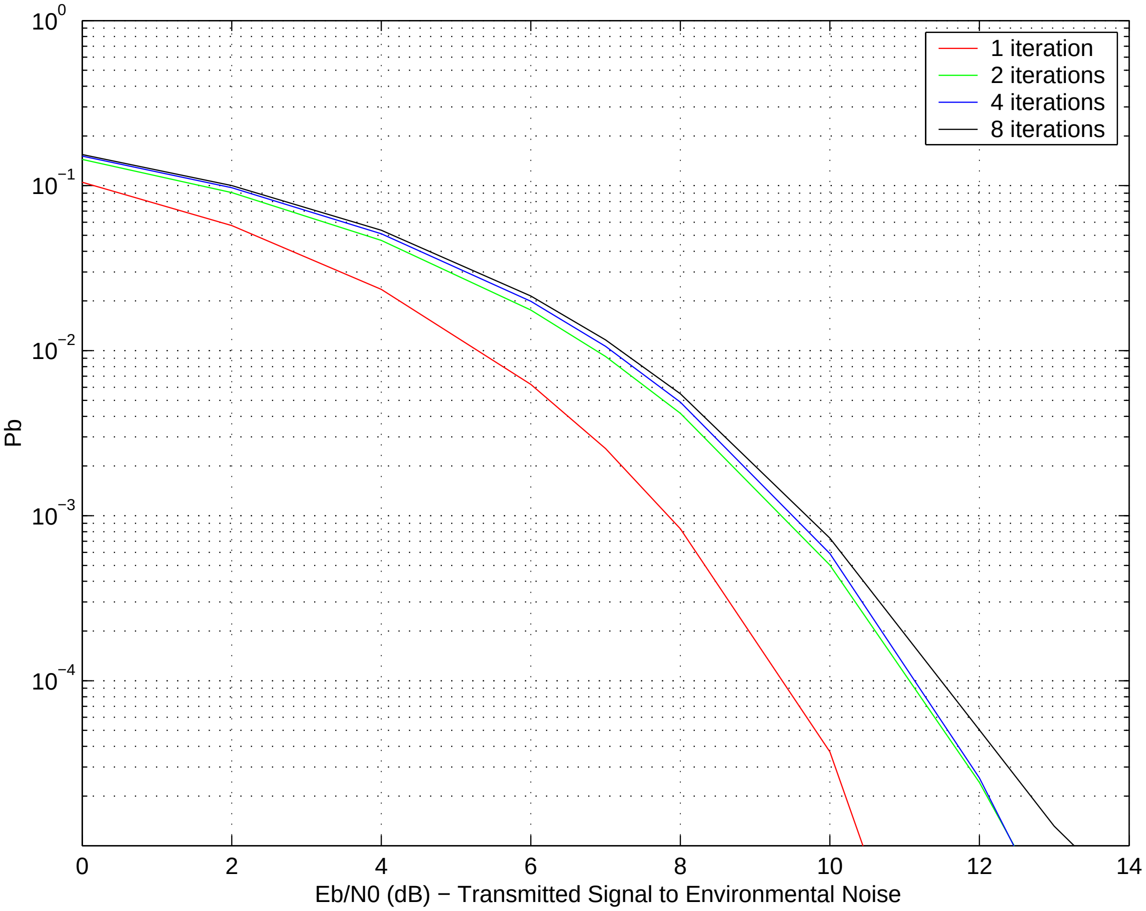

Figure 3.12: Complexity: 16 Equal Power Users, Not Coded

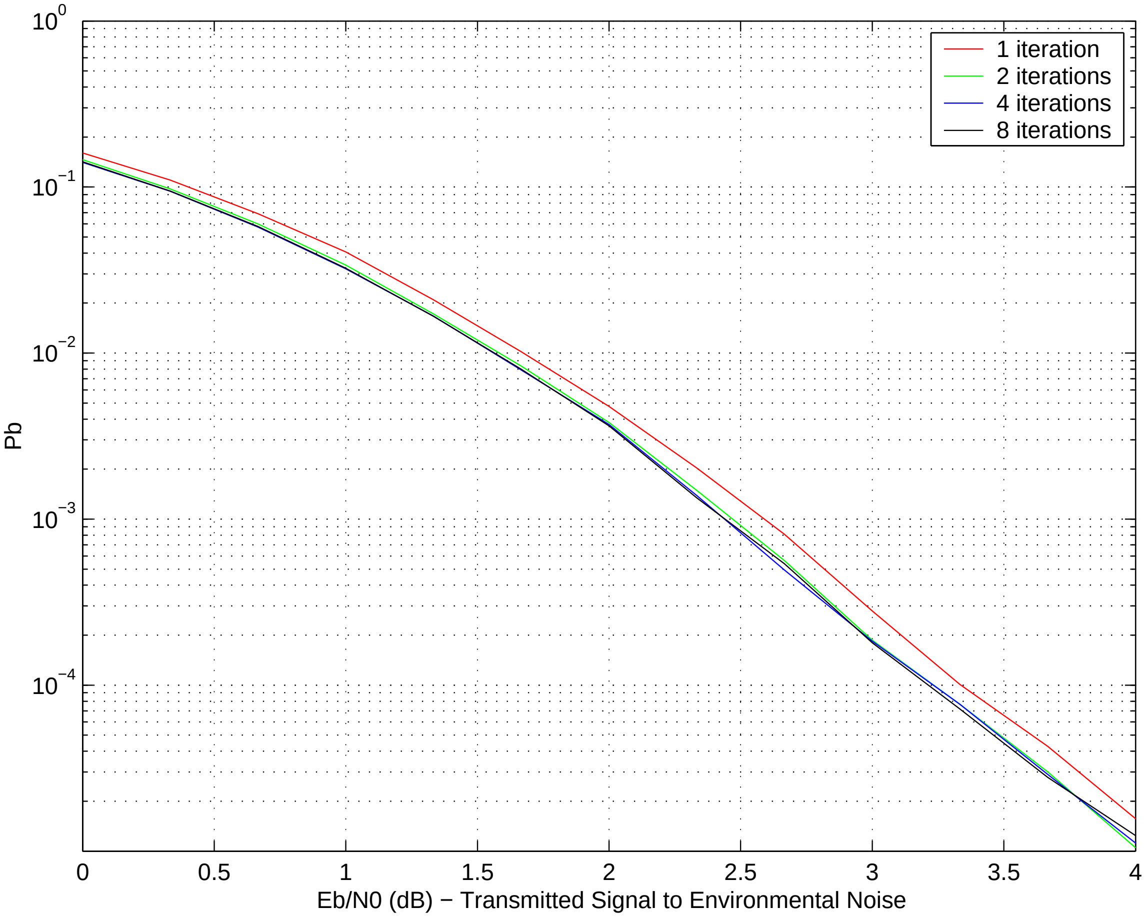

Figure 3.13: Complexity: 8 Equal Power Users, Coded

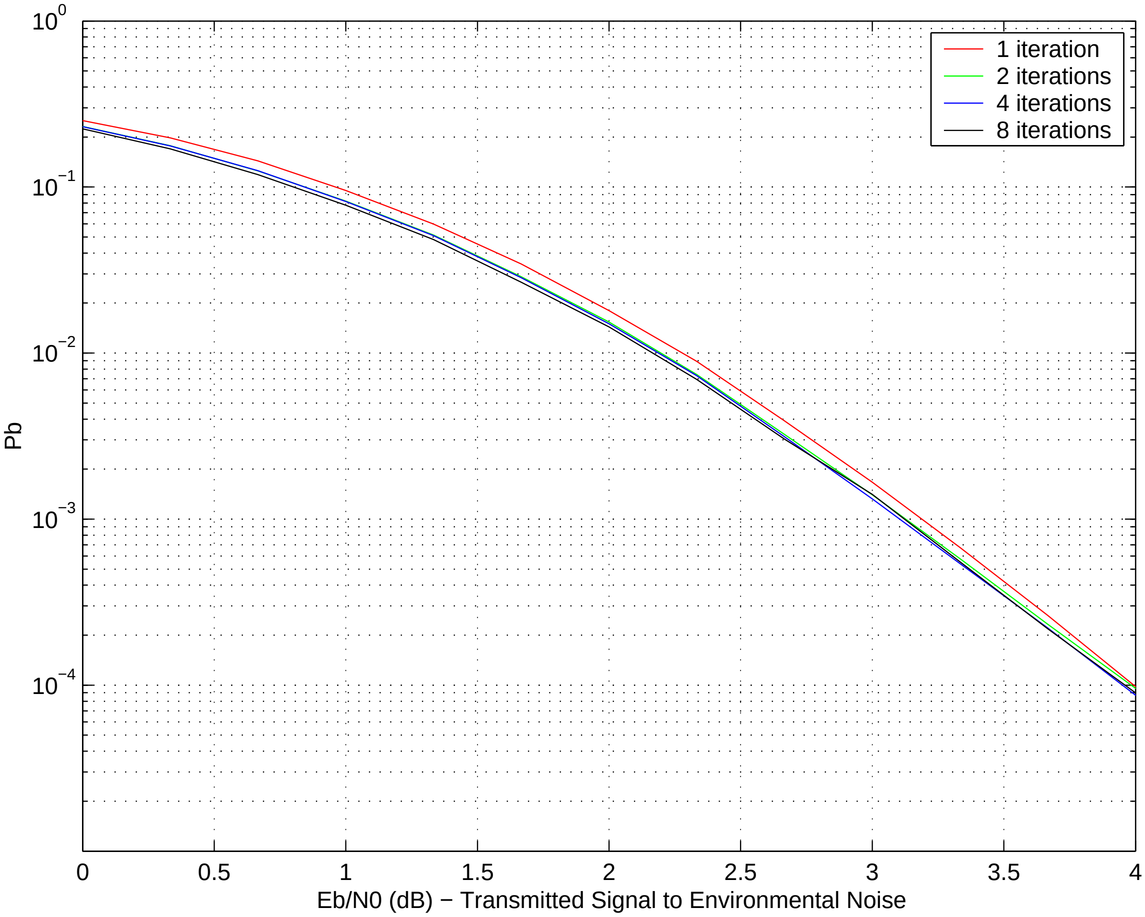

Figure 3.14: Complexity: 16 Equal Power Users, Coded

The complexity of the proposed inequality constrained method compares favorably with that in [10]. Usually, only three Cholesky factorizations are required (two for the loop, plus an additional factorization to obtain the solution $\mathbf{b}*$). Each factorization is performed on a square $L \times L$ matrix. In [10], a singular-value decomposition of an $M \times M$ matrix must be performed followed by a bisection search where each iteration uses an $O(M^2)$ procedure.

Chapter 4

Conclusions

Several questions remain regarding the performance of the methods proposed in this thesis. The results provided in the previous section give a reasonably accurate impression of the performance that can be achieved. However, as mentioned earlier, they are not thoroughly optimized. With extensive simulation, optimized combinations of iteration limits and energy constraints could be found and stored in a table to give improved performance along the entire BER curve. The limits of performance that can be achieved in this manner is still unknown. Also unknown is whether an improved model is required. The ad hoc approach used in this paper resulted in good performance, but does not explicitly encapsulate the concept of BER performance. The potential for better performance from better models remains.

Despite these uncertainties, a novel method has been demonstrated in this thesis which, for practical systems,

- delivers gains of 0.75dB to 4dB over existing unconstrained detection methods in coded equal power environments

- delivers gains of 0.25dB to 0.5dB over existing unconstrained detection methods in coded near/far environments
- delivers gains of 0.25db to 0.5dB over existing constrained precoding methods in coded equal power environments
- delivers gains of 0.1dB to 0.25dB over existing constrained precoding methods in coded near/far environments
- requires only 2 to 4 linear system solutions using fast Cholesky factorizations to achieve those gains

What seems clear is that energy constraints will be an important component of any future methods. Energy is too fundamental to the performance of wireless communication systems to allow to drift unchecked, and, as long as energy constraints remain, constrained quadratic programming will remain an important technique in the system engineer's toolkit.

Chapter 5

Further Research

In the previous chapter, all results for the optimizing precoder were tuned for optimal performance at a BER of 10^{-3}. Improved performance may be possible at lower error rates if the r parameter is tuned individually for each of those error rates. It is worth discovering the upper limits of performance that can be achieved this way. It also remains to be seen how the specific channel coding scheme affects performance. Performance of the optimizing precoder with block codes or turbo codes should be investigated.

The simulation study was performed in a highly idealized environment with full synchronism and no multipath. Performance of the optimizing precoder in asynchronous environments is possible, and may yield even greater gains. Multipath environments should also be investigated as multipath resistance may be affected by the optimization. There is potential to apply the proposed methods to RAKE receivers as well.

The optimizing precoder attempts to minimize MMSE at the receiver with a constraint on energy. More promising are quadratic programs which minimize energy with limits on the MAI to specific users. This can be particularly useful in near/far environments. Indeed, many alternative formulations exist with other quadratic optimization methods.

References

[1] J. Nocedal and S. J. Wright, *Numerical Optimization.* New York : Springer, 1999.

[2] M. K. Kozlov, S. P. Tarasov, and L. G. Hacǐjan, “Polynomial solvability of convex quadratic programming,” *Soviet Mathematics Doklady*, 20, 1979, no. 5, pp. 1108-1111.

[3] K. Murty and S. Kabadi, “Some NP-complete problems in quadratic and nonlinear programming, *Mathematical Programming*, 19 (1987), pp. 200-212.

[4] Y. Ye, “On affine scaling algorithms for nonconvex quadratic programming,” *Mathematical Programming*, 56, 1992, pp. 285-300.

[5] S. Miri, E. Hons, and A. Khandani, “On optimizing the combined source and channel coding of a discrete communication system”, *submited for publication.*

[6] R. Rockafellar, “Lagrange multipliers and optimality,” *SIAM Review*, 35 (1993), pp. 183-238.

[7] N. J. Higham, “FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation,” *ACM Trans. Math. Soft.*, 14 (1988), pp. 381-396.

[8] S. Haykin, *Digital Communications*. Toronto: John Wiley and Cons, 1988.

[9] S. Verdú, “Optimal Multi-User Signal Detection,” Ph.D. Thesis, University of Illinois at Urbana-Champaign, Coordinated Science Laboratory, Urbana, U.S.A., September 1984.

[10] B. R. Vojǒić and W. M. Jang, “Transmitter precoding in synchronous multiuser communications,” *IEEE Trans. Comm.*, 46 (October 1998), pp. 1346-1355.

[11] W. M. Jang, B. R. Vojǒić and R. L. Pickholtz, “Joint Transmitter-Recevier Optimization in Synchronous Multiuser Communication over Multipath Channels”, *IEEE Trans. Comm.*, 46 (February 1998), pp. 269-278.

[12] M. Brandt-Pearce and A. Dharap, “Transmitter-Based Multiuser Interference Rejection for the Down-Link of a Wireless CDMA System in a Multipath Environment,” *IEEE J. Select. Areas Comm.*, 18 (March 2000), pp. 407-417.

[13] S. Verdú, “Minimum Probability of Error for Asynchronous Gaussian Multiple-Access Channels,” *IEEE Trans. Inform.* Theory, 32 (January 1986), pp. 85-96.

[14] S. Verdú, “Computational Complexity of Optimum Multiuser Detection,” *Algorithmica*, v. 4, 1989, pp. 303-312.

[15] R. Lupas and S. Verdú, "Linear Multiuser Detectors for Synchronous Code-Division Multiple-Access Channels," *IEEE Trans. Inform.* Theory, 35 (January 1989), pp. 123-136.

[16] A. Duel-Hallen, J. Holtzman, and Z. Zvonar, "Multiuser Detection for CDMA Systems," *IEEE Personal Comm.*, 2 (April 1995), pp. 46-58.

[17] S. Verdú, *Multiuser Detection.* New York: Cambridge University Press, 1998.

[18] H. D. Schotten, H. Elders-Boll, and A. Busboom, "Optimization of spreading-sequences for DS-CDMA systems and frequency selective fading channels," *Proceedings of ISSSTA'98 International Symposium on Spread Spectrum Techniques and Applications*, v. 1, 1998, pp. 33-37.

[19] P. J. E. Jeszensky and G. Stolfi, "CDMA systems sequences optimization by simulated annealing," *Proceedings of ISSSTA'98 International Symposium on Spread Spectrum Techniques and Applications*, v. 1, 1998, pp. 38-40.

[20] K. L. Cheah, S. W. Oh, and K. H. Li, "Efficient performance analysis of asynchronous cellular CDMA over Rayleigh-fading channels," *IEEE Communications Letters* 1 (May) 1997, pp. 71-73.

[21] S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inform.* Theory, 45 (March 1999), pp. 622-640.

[22] A. Viterbi, *CDMA: Principles of Spread Spectrum Communications.* Reading, MA: Addison-Wesley, 1995.